Uncertainty Quantification for Deep Context-Aware Mobile Activity Recognition and Unknown Context Discovery

Zepeng Huo[1] **Arash Pakbin**[1] **Xiaohan Chen**[1] **Nathan Hurley**[1] **Ye Yuan**[1]
Xiaoning Qian[1] **Zhangyang Wang**[1] **Shuai Huang**[2] **Bobak J. Mortazavi**[1]
Texas A&M University[1] University of Washington[2]

Abstract

Activity recognition in wearable computing faces two key challenges: i) activity characteristics may be context-dependent and change under different contexts or situations; ii) unknown contexts and activities may occur from time to time, requiring flexibility and adaptability of the algorithm. We develop a context-aware mixture of deep models termed the α-β network coupled with uncertainty quantification (UQ) based upon maximum entropy to enhance human activity recognition performance. We improve accuracy and F score by 10% by identifying high-level contexts in a data-driven way to guide model development. In order to ensure training stability, we have used a clustering-based pre-training in both public and in-house datasets, demonstrating improved accuracy through unknown context discovery.

1 Introduction

In wearable computing, context-awareness helps recognize activities based on sensor measurements under different situations. Context can be defined as "any information that can be used to characterize the situation" [Sezer et al., 2018] or to improve recognition [Dey, 2001]. Context-aware systems have previously been used in many applications, including activity recognition [Riboni and Bettini, 2011], online, personalized and adaptive activity classification [Xu et al., 2016], and healthcare applications [Andreu-Perez et al., 2015; Spiegel et al., 2014]. The definition of context heavily relies on domain knowledge, such as a user's tasks (e.g., spontaneous activity, engaged tasks) or a user's social environment (e.g., co-location of others, group dynamics), etc. However, in practice, pre-defined contexts may not always be available, or definitions of contexts may change in different environments. Additionally, new unknown contexts may emerge over time. For these reasons, there is a general data insufficiency and lack of contextual information to develop accurate context-aware activity recognition systems that could adapt to these unknown contexts. While existing works are not adequate to address the challenges such as a lack of context information in data while training the model, or the emergence of unknown contexts when the model is applied [Xu et al., 2014; Steven Eyobu and Han, 2018], we propose a data-driven approach with context-awareness capability to achieve better activity recognition performance. Specifically, we develop an integrative α-β framework to simultaneously learn unknown contexts and the distribution of each user's specific activity likelihood within each context. In this framework, the α network is the context detector to learn a distribution over contexts as a mixture of weights, and the β network models activity recognition for context-specific sensor measurements. For example, given the sensor reading data from a user, the α network detects the context by generating a distribution over different contexts; then, each context has a dedicated β network that outputs a distribution over different activities.

We further extend our model with the ability to explore new unknown contexts by equipping the α-β network with uncertainty quantification (UQ) based on the maximum entropy learning (MEL) principal. MEL identifies the distribution of the parameters of a statistical model that bears the maximum uncertainty, rather than one single best model, as a principle to achieve robustness in prediction and modeling. The prediction model could refuse to predict on given data if the uncertainty for making a prediction on this data is higher than a threshold. This method adapts to data and effectively discovers unknown contexts with the UQ. This work allows for models trained in labo-

Proceedings of the 23[rd]International Conference on Artificial Intelligence and Statistics (AISTATS) 2020, Palermo, Italy. PMLR: Volume 108.

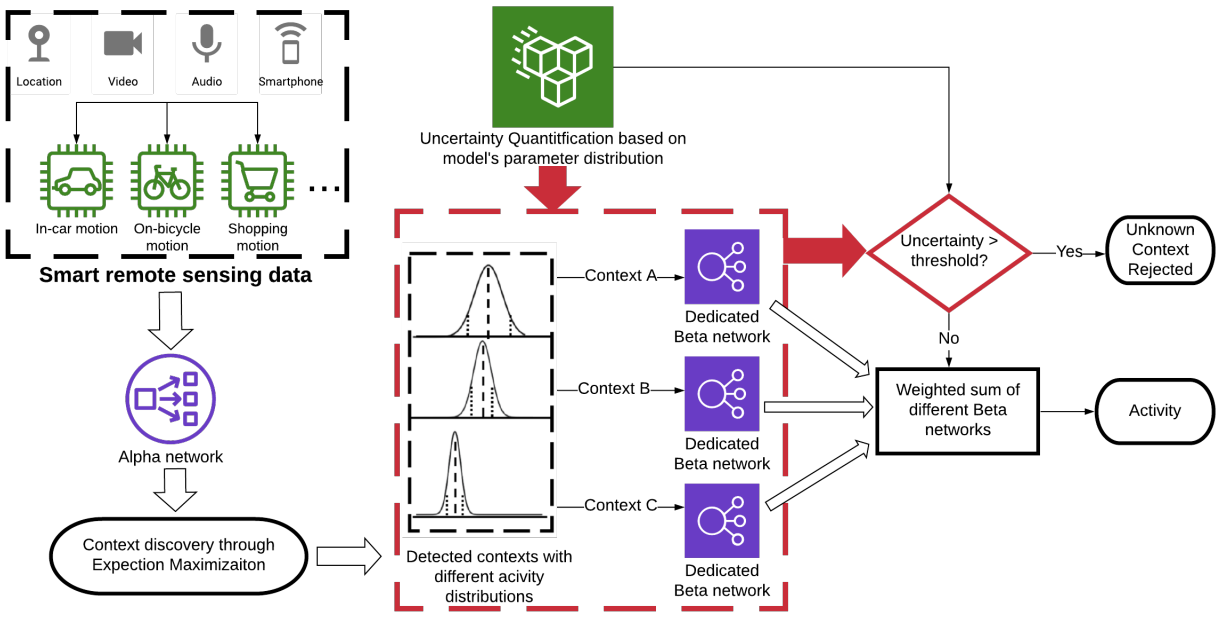


Figure 1: A conceptual overview of UQ integrated with the α-β network

ratory settings to extend to natural environments for monitoring behaviors and performances of users, as in Figure 1.

Our contributions are as follows. In this paper, we propose a context-aware model for activity recognition. The context and activity are simultaneously modeled by dedicated networks. For unknown contexts, an overarching UQ method is applied to all the model parameters. This provides robustness in testing new context that our model can uniquely offer, beyond a traditional activity recognition technique. Finally, we demonstrate these findings in both a publicly-available benchmark dataset and an in-house dataset we collected to identify confounded versions of human motion, and make this data available for public use.

2 Related Work

Context-Awareness: Mixture of Experts Model In many applications of machine learning, heterogeneous data can be divided into smaller homogeneous groups that can be modeled more accurately. Context-awareness, as an example, plays an important role in improving the performance of activity recognition systems [Ravi et al., 2005; Wang et al., 2019]. In Mixture of Experts (MoE) models [Jordan and Jacobs, 1994] such group-level clustering and modeling comes as a single training step, rather than splitting data a prior then building models. Elements are clustered based upon their relationship and the next level modeling accuracy. MoEs have successfully served different applications from classification and regression tasks [Yuksel et al., 2012][Miech et al., 2017] to phenotyping in medical datasets [Courbariaux et al., 2018]. In [Jordan and Jacobs, 1994][Yuksel et al., 2012][Jordan and Xu, 1995], authors provided formulations of probabilities of observed given different experts. The MoE structure consists of two components: A gate and several experts. The gate is often modeled by Gaussian Mixture Models (GMM) [Yuan and Neubauer, 2009][Sharma et al., 2019] and neural networks [Lima et al., 2007], while expert modeling is more dependant on the application including SVM [Lima et al., 2007] [Cao, 2003]. This work used neural networks for both.

Different methods have been used in the literature in order to determine the number of experts for these models [Yuksel et al., 2012]: growing models where the experts with the worst performance decompose into a MoE themselves [Shazeer et al., 2017] [Aljundi et al., 2017], pruning models where they start with a large number of experts and reduce the number by combining/removing experts [Jacobs et al., 1997], and exhaustive search in cases where tree topologies are not overly complex [Bishop and Svenskn, 2002].

Uncertainty Quantification: Previous Studies UQ has been critical for robust learning under different contexts (known or unknown) in mobile activity recognition [Ardywibowo et al., 2019], healthcare [Samareh and Huang, 2019; Meghdadi et al., 2017], signal processing [Reynders et al., 2016], and manufacturing [Nannapaneni and Mahadevan, 2014]. Existing models with Gaussian process [Ardywibowo et al.,

2019] or Bayesian approximation methods [Gal and Ghahramani, 2016] either rely on the assumption that variables in a system can be characterized by explicit probabilistic relationships (e.g., Bayesian models) or rely on generating one best model in the learning algorithm. However, measuring the predictive uncertainty for deep neural networks remains a challenging problem, closely related to the problem of detecting test samples that are drawn sufficiently far away from the distribution of training samples. For example, Malinin & Gales proposed a framework called Prior Networks (PNs) for modeling predictive uncertainty that explicitly modeled distributional uncertainty [Malinin and Gales, 2018], Hendrycks & Gimpel proposed a framework that utilized probabilities from softmax distributions and detected out-of-distribution examples, by introducing confidence scores based on density estimators [Hendrycks and Gimpel, 2016], later improved by processing the input and output of DNNs in Liang et al. [2017]. Lee et al. proposed a method for detecting abnormal test samples by including both out-of-distribution and adversarial samples, to obtain the class conditional Gaussian distributions by introducing confidence scores based on the Mahalanobis distance [Lee et al., 2018]. Our method is different from these works as we aim to learn a distribution of deep models rather than one single best one.

3 Methods

Contextual information can help model similar activities together, resulting in an improvement of activity recognition performance by reducing the search space of activities to recognize given a set of features. Our proposed α-β network integrates activity recognition with unsupervised context detection, as detailed in Section 3.1. As context can vary from person to person, and change over time for the same person, in Section 3.2, we present maximum entropy learning (MEL) based UQ in the α-β network to discover unknown contexts when needed.

3.1 Context-Awareness Processing

In this work we develop a mixture of CNNs, the α-β network, where each mixture component is dedicated to one specific context. There are two types of networks: α and β. Given the sensor data, the α network detects context by generating a probability distribution over all known contexts. Each context has a dedicated β network that outputs a probability distribution over different activities. Our activity recognition problem features a latent context variable and can be formulated as:

$$\begin{aligned}\log p(ACTIVITY|\mathbf{X},\theta) &= \sum_{i=1}^{N}\log p(activity_i|\mathbf{x}_i,\theta)\\ &= \sum_{i=1}^{N}\log\sum_{c=1}^{N_c} p(activity_i|c_i=c,\mathbf{x}_i,\theta)p(c_i=c|\mathbf{x}_i,\theta),\end{aligned} \tag{1}$$

where θ denotes the mixture component parameters, N denotes the number of data samples, and N_c denotes the number of expected clusters (contexts) to which each data point may belong. Our objective is to maximize Eq. (1) with respect to θ. As shown by [Dempster et al., 1977], the log-likelihood has a lower bound, first formulated in [Solis et al., 2019]:

$$\log p(ACTIVITY|\mathbf{X},\theta) \geq \sum_{n=1}^{N}\sum_{c=1}^{N_c} q(c_i=c)\log\frac{p(activity_i|c_i=c,\mathbf{x}_i,\theta).p(c_i=c)}{q(c_i=c)},$$

where $q(\cdot)$ is the distribution over different contexts and $p(\cdot|context,x,\theta)$ is the distribution over activities given the context and the input. $q(\cdot)$ is modeled using the context-detecting α network, and each $p(\cdot|context,x,\theta)$ is modeled using a β network (each context has its own β network). While [Solis et al., 2019] used a supervised technique to define contexts as specific locations, this work provides for an unsupervised exploration of context. Additionally, in our implementation, we use a network for context recognition (α network) and a dedicated network for each context (β network) whereas [Solis et al., 2019] used only two networks regardless of the number of contexts. Following the EM algorithm, the lower bound in Eq. (2) can be maximized. Specifically, the loss, the negative of the lower bound, is minimized. In the E-step, $q(\cdot)$ is optimized which translates to optimizing α network while freezing β networks. In the M-step, θ (model parameters) need to be optimized which translates to optimizing β networks while freezing α network. The EM training alternates iterations of training either α network or β networks while keeping the other(s) fixed. It should be noted that no labeled contextual data is used in the training process for this α-β network.

3.2 Unknown Context Discovery

The α network enables context detection; however, in practice, contexts may change over time, or may not always be pre-defined. It is possible to improve context-aware systems by detecting the uncertainty of possible unknown contexts as a result of potential distribution mismatch between known and unknown contexts. To identify unknown contexts, we combine the feature extraction power of deep learning with the learning power of MEL to define a probabilistic mechanism for unknown context discovery.

While many of the current works focus on revising general deep models with a probabilistic evaluation of their model or prediction, here we have a different aim: We modify the α-β network to be adaptive to changing contexts hidden in data. We relax our expectation of identifying one single optimal model of the α-β network; rather, we consider solving for a full distribution over multiple models. The intuition is that many different models might generate relatively similar performance, so it would be better to estimate a distribution over parameters $p(\mathbf{w})$, from the output layer of the α networks that detects context. This aligns with the basic principal of MEL [Sensoy et al., 2018; Bauer et al., 2019]. Therefore, we equip the α-β network with UQ capacity based on the MEL principal by identifying the distribution of the parameters of a statistical model that bears the maximum uncertainty.

3.2.1 Uncertainty Quantification via Minimizing Relative Entropy

To learn the distribution of the α-β network parameters that encode maximum uncertainty, we employ the MEL formulation. There are two steps. First, we create constraints that encode information from the data. For example, for each sample we derive a loss function such that the expected prediction on this sample over all the possible model parameters matches the observed outcome on this sample, as in traditional ML. Second, on the top of this constraint structure, the learning objective of MEL is to learn the distribution of the model parameters with the maximal entropy in terms of the parameter posterior distribution. Thus, unlike traditional machine learning methods that estimate a single optimal setting of the parameter, MEL considers a more general problem of these methods by solving for a full distribution over multiple $p(\mathbf{w})$ values.

3.2.2 Analytical Details of MEL

To further illustrate this distribution approach, note that our context detection problem is also a classification problem where the response variable is denoted by y taking values for different contexts. Let $\mathbf{x}_n = [\mathbf{x}_1, \ldots, \mathbf{x}_n]$ be an input feature vector as an aggregate of all the measures from sensors for each window, and let $\mathcal{D}(\mathbf{x}_n|\mathbf{w})$ be the discriminant function parameterized by $\mathbf{w}$ implemented in the α network. Traditional learning machines such as the max-margin methods estimate the optimal $\hat{\mathbf{w}}$ that minimizes the classification error in predicting the labels of training examples as:

$$\hat{y} = sign\mathcal{D}(\mathbf{x}_n|\mathbf{w}). \tag{2}$$

Based on this line of thought, we can classify margin as $y_n\mathcal{D}(\mathbf{x}_n, \mathbf{w})$, and learn the optimal parameter setting $\mathbf{w}$ by the empirical loss and the regularization penalty as:

$$\begin{aligned} &\min_{(\mathbf{w},\gamma_n)} R(\mathbf{w}) + \sum_n L(\gamma_n) \\ &\text{s.t.} \quad y_n\mathcal{D}(\mathbf{x}_n \mid \mathbf{w}) - \gamma_n \geq \mathbf{0}, \quad \forall n. \end{aligned} \tag{3}$$

where $L(\gamma_n)$ is the loss function, a non-increasing and convex function of the margin, and $R(\mathbf{w})$ is the regularization penalty. Given $p(\mathbf{w})$, we can recast (3) as an integration where the classification constraints will also be applied in an expected sense. Instead of considering an expectation of the regularization penalty functions, we can apply a canonical penalty function for distributions, the negative entropy; minimizing the negative entropy is equivalent to maximizing the entropy. Hence, we use the Shannon entropy defined as $H(p(\mathbf{w})) = -\int p(\mathbf{w}) \log p(\mathbf{w}) d\mathbf{w}$. This gives us the following objective function to learn the distribution $p(\mathbf{w})$ over the parameters $\mathbf{w}$:

$$\begin{aligned} &\min_{p(\mathbf{w})} H(p(\mathbf{w})) \\ &\text{s.t.} \quad \int p(\mathbf{w})[y_n\mathcal{D}(\mathbf{x}_n, \mathbf{w}) - \gamma_n] d\mathbf{w} \geq \mathbf{0}, \quad \forall n. \end{aligned} \tag{4}$$

As a result, MEL no longer finds a fixed set of the parameters, but a distribution over them. Learning such a distribution of model parameters does not rely on assumptions on the model's mathematical form. It also does not rely on knowing a particular distribution as is needed in Bayesian learning frameworks. Therefore, MEL is more flexible than typical Bayesian learning methods [Sun et al., 2018; Zhu and Wang, 2018] to characterize uncertainties associated with complex models such as the α-β network here. To solve the MEL formulation (4), we could derive a Lagrangian $J(p)$, and take the derivatives with respect to $\mathbf{w}$ and set them to 0. To do that we first need to calculate the unconditional maximum of the problem (4) plus the constraints added with some multiplying factors (the Lagrange multipliers), which give the probabilities in a functional form with the Lagrange multipliers as parameters. Our UQ approach compares the uncertainty with a threshold to see whether a given sample should be detected as belonging to a new context or not. We defined a classification with rejection option as $\hat{y}_i^{Rej}$, where if a sample is rejected $\hat{y}_i^{Rej} = 0$, and if it is accepted $\hat{y}_i = \hat{y}$, where $\hat{y}$ corresponds to the classification of the ith sample. Note that, a sample is rejected when $p(\mathbf{w}|x_i) < \epsilon$, and ϵ is chosen through cross-validation.

The distribution of parameters forms a quantitative evaluation of model uncertainty, which could be further used in subsequent decision making by using probability laws to track the uncertainty propagation process. In this paper, we create a rejection option, a flexibility enabled by the UQ capacity. The rejection

option allows for the prediction model to refuse to generate a prediction if the uncertainty is higher than a given threshold. This is typically solved by estimating the class conditional probabilities and rejecting the samples that have lower posterior probability of class.

4 Experiments

In this section, we discuss experiments with our deep context-aware mixture of experts for activity detection coupled with maximum entropy based uncertainty quantification. We first introduce the datasets, and then the detailed implementation of our models. Then, we show various competitive baselines to demonstrate that our pipeline performs the best as evaluated by several different metrics.

4.1 Datasets

UCI data. We used the UCI OPPORTUNITY dataset [Roggen et al., 2010] for context-aware human activity recognition. The dataset contained 18 different activities performed in five different contexts and sensed by 72 different sensors. Each of the 18 activities had one of the five contexts, but not all contexts contained every activity. Therefore, the UCI OPPORTUNITY dataset provided a realistic capture of a situation where not all of the human activities occurred with an equal likelihood in all contexts. The UCI OPPORTUNITY dataset has seven levels of hierarchical labels. Higher level labels described details such as subject posture, while lower level labels described the hand movements or interactions with other subjects. In this study, we chose a higher level label (e.g., cleaning time) as the context and a lower level (e.g., opening a door) label as the activity. We used all the body-worn sensors which included seven inertial measurement units (IMUs) and twelve 3D acceleration sensors. Five IMUs were on the upper body while two were on user's shoes. Accelerometers were on the upper body, hip, and leg, which translated to 133 columns in the raw dataset.

In-house data. To generate a more realistic experimental setting, we have collected our data to detect different types of human motion that may be confounded by environments factors. The dataset is made publicly available, serving as extra part of our contribution in this paper. The motivations are three fold: 1) instead of collecting the data in a strict laboratory setting, we have loosened all the rules for subjects, including their choice of rest time and the pace of activity. 2) we have devised a set of much more realistic contexts to be detected. Those contexts can cover almost all the real world setting a subject might face. 3) we have collected a much nosier dataset compared to previous ones, with respect to the randomness we imposed on the data collection, such as a rare activity as walking with one shoe off in both outdoor (lawn) and indoor (hardwood covered by carpet). As a result we have 3 contexts with corresponding movements (1) outdoors: Crawling, Jogging, Riding Bike, Sprinting, Walking, Walking with One Shoe (simulated limping), Walking with Weight in Arms, Walking with Weight on Back. (2) Movements that happen indoors: Escalator Up, Escalator Down, Elevator Up, Elevator Down, Lying Down, Sitting, Stairs Down, Stairs Up, Standing, Walking, Walk with One Shoe (simulated limping), Cooking, Dancing, Eating, Reading, Sleeping, Talking on phone, Talking to Another Person, Using PC, and (3) movements that happen outdoors but in vehicles: Driving Car, Riding Car, Riding Bus, Reading, Device Usage. The movements, however, are not unique to each context, and the position of the phone may change through usage. We collected data on 20 people, each doing 32 activities while having 3 phones, one in hand, one in the pocket and one in the backpack, as those are the common places for phone positions. The applications used on the phones for data collection was *Readisens* [Lockheed Martin, 2019]. The sensors used are: acceleration (in three axes), altitude, compass, gyroscope (in three axes), GPS information (latitude, longitude), screen time information, and phone speed. This data contains contextual information which is independent of the locations of the phone. Participants were given minimal instruction for the execution of the activities in order to allow for individual variation. Participants were also wearing clothes of choice and the order of the activities were randomized. This study was reviewed and approved by the Texas A&M Institutional Review Board (IRB # 2018-210D). The data has been made available for public use.[1]

4.2 Implementation

In UCI dataset, 19 different preprocessed sensors were fed into the network. Time series data were divided into non-overlapping segments of 1 second (30 samples). In each window, the features of sensors are concatenated. We used a five-fold cross-validation to evaluate our models with testing accuracy and micro F score as performance metrics. In our experiments, we used 3 convolutional layers followed by 3 fully connected layers for both α and β networks. Note that these two networks' architectures were different in terms of the number of neurons in the output layer. Networks were trained using stochastic gradient descent with an initial learning rate of 0.001 and a momentum of 0.9, which provided the best results in

[1] https://github.tamu.edu/guangzhou92/RealActivity

cross-validation.

Implementation of MEL. For maximum entropy classifier, we have the input from the parameter distribution derived from the α network. This network outputs $\hat{y}_i^{Rej}$ as a probability of rejection, which is compared against a threshold that is derived from cross-validation to further guide the β network for fine-grained activity detection under specific context. We defined a classification with rejection, where if a sample was rejected (if the uncertainty was higher than a specific threshold), the prediction model refused to generate a prediction by setting the predicted context to zero.

In the UCI OPPORTUNITY dataset we had five contexts: relaxing, coffee time, early morning, clean-up, and meal time. To test our unknown context discovery, we adopted a rotating strategy. In this strategy, we removed one context and its corresponding data from the training dataset at each rotation, and trained an $\alpha - \beta$ network only on the remaining known contexts. In this case, one context was assumed to be unknown and was treated as a hold-out to be used for unknown context discovery assessment. The final evaluation was conducted through comparison of activities that were sampled from both known and unknown contexts, to demonstrate the model's ability to distinguish the unknown context.

Pre-training. Initialization is an essential step for both the optimization and for the training of the neural networks. Without proper initialization, the model collapses into selecting only one specific β network while eliminating the contribution of others, as is observed in many other mixture network modeling. We have added pre-training to solve this problem and have compared it with regularization. We have shown that it is much more effective in terms of accuracy. Pre-training is an important stage in the α-β network training. The model could approach the base model of a single neural network classifier without proper pre-training because of the large gradients at the beginning of the training stage; large gradients cause the selector to saturate and select only one β network. Therefore, the full capabilities of the α-β network can only be achieved using proper pre-training, which proves useful in finding subgroups of data as well as in yielding a better performance. We used the idea presented in [Guérin et al., 2017] to cluster the activity data. In detail, a base network was trained with sensor readings as input and activities as output. Next, the CNN segment of the network was used to embed the sensor readings into the features which have proven to be descriptive of the input data [Sharif Razavian et al., 2014; Donahue et al., 2014]. Subsequently, we used K-means to cluster the input into a fixed number of clusters. Finally, we trained our α network to learn the mapping between sensor readings of activity to clusters.

Table 1: UQ results VS. baseline for unknown context discovery where contexts are 1 = Relaxing, 2 = Coffee time, 3 = Early morning, 4 = Clean up, 5 = Sandwich time

Performance measure	Context number that was removed each rotation				
	1	2	3	4	5
UQ					
Sensitivity	0.63	0.71	0.75	0.62	0.73
Specificity	0.72	0.72	0.76	0.79	0.82
Testing accuracy	0.67	0.71	0.75	0.69	0.77
F-score	0.67	0.72	0.77	0.70	0.78
Baseline					
Sensitivity	0.26	0.34	0.17	0.19	0.26
Specificity	0.75	0.79	0.81	0.80	0.80
Testing accuracy	0.66	0.72	0.64	0.64	0.69
F-score	0.26	0.34	0.17	0.20	0.27

4.3 Baseline

We compare our model against several baselines. The first baseline is a single β network, which is a component of the α-β network. Another baseline, for each specific number of contexts, is a β network which has wider hidden layers in order to have the number of parameters equal to the α-β output (similar size). In other words, for a α-β network with the number of clusters equal to k, given that α and β networks have the same size, the baseline is a β network with hidden layers $k+1$ times as large to have roughly the same capacity. Finally, in order to demonstrate the performance of our UQ method, we design a similar probabilistic baseline, logistic regression, to output a rejection likelihood and we compare it with our method to show the better unknown context discovery performance from our UQ pipeline. Baseline doesn't impose extra regularization, and all parameters were selected through cross-validation.

4.4 Result of UCI dataset: UQ vs. Baseline

For the performance of UQ against baseline, results are shown in Table 1. As can be seen, the proposed UQ method detects unknown contexts better than baseline in all evaluations.

Figure 2 presents a main result of UQ in this experiment from the UCI OPPORTUNITY dataset. Here, the distribution of predicted probabilities of the removed context (the red histogram) is shown with the distribution for all the other contexts combined (the blue histogram). The UQ algorithm is able to detect the uncertainty for the unknown context, as the un-

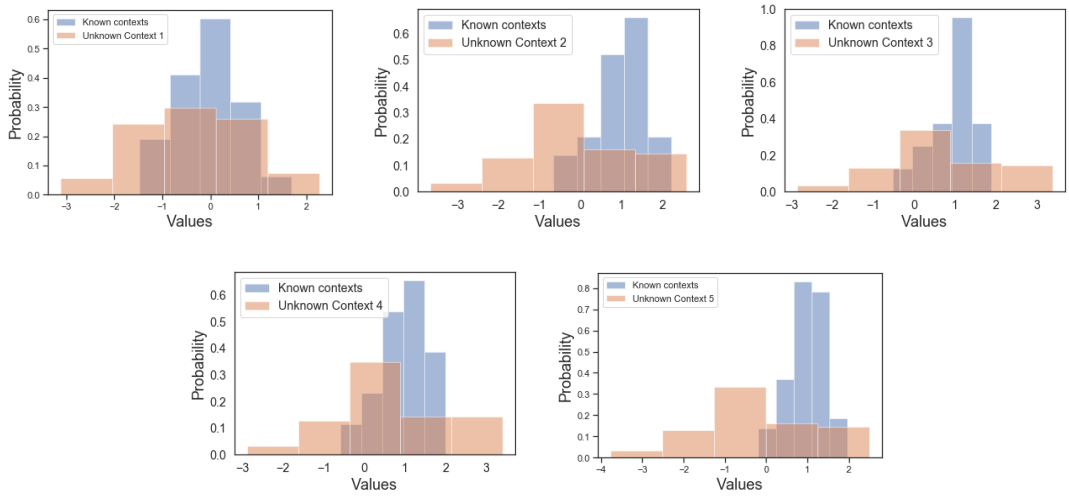


Figure 2: Predicted probability distributions when a context is removed v.s the aggregate of all known contexts within each rotation.

Table 2: Accuracy (F score) for the α-β network and baseline which is a β network with the same number of parameters for each specific number of clusters. For 1 cluster both the models are the same.

Clusters	2	3	4
α-β	**0.89(0.90)**	**0.89(0.90)**	**0.91(0.91)**
baseline	0.81(0.81)	0.84(0.84)	0.84(0.83)
Clusters	5	6	7
α-β	**0.92(0.91)**	**0.91(0.92)**	**0.96(0.96)**
baseline	0.83(0.82)	0.86(0.86)	0.85(0.85)
Clusters	8	9	10
α-β	**0.96(0.96)**	**0.96(0.96)**	**0.95(0.95)**
baseline	0.84(0.84)	0.86(0.86)	0.86(0.86)

known context usually leads to smaller probabilities in comparison with the distribution of the known contexts.

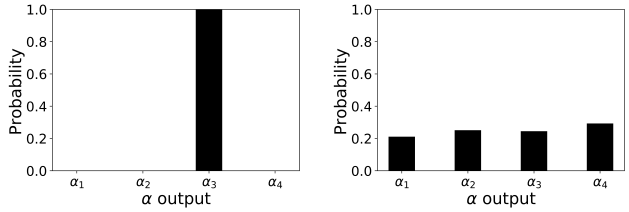
4.5 Results of UCI dataset: α-β Network vs. Baseline

Table 2 compares the testing accuracy of the α-β network and the baseline (a single network that is equivalent to a single β network) for predicting labels in the UCI OPPORTUNITY dataset. The testing accuracies are averaged among all their corresponding bootstraps (bootstrapped 5 times). Table 2 shows that the mixture of the context-specific neural networks improved on the accuracy of the baseline from 86% to 96% when using nine contexts. Thus, our context-aware α-β network was able to find subgroups in the data in an unsupervised manner with different numbers of contexts This fact was reflected in the accuracy boost due to the subgroup modeling by different β networks.

The α-β network is equipped with pre-training. The comparison in Figure 3 shows that pre-training is extremely important, as the model without pre-training results in selection of only one network, leading to model collapse. The effect of pre-training can also be seen in the network performance. Figure 4 shows that pre-training is much more effective than regularization since regularization just penalizes single network usage but does not use any knowledge about the data in the process. We took the best number of contexts in terms of accuracy and F-score, 9, and tried to train networks with the same number of contexts but without pre-training and by using only regularization with different regularization coefficients. Without proper initialization, the α-β network drops from 96% to 87% in accuracy and 0.96 to 0.87 in F-score. This is 4% less than the performance of an identical α-β network with proper initialization in Table 2. This is roughly equal to the performance achieved by a single β network (refer to baseline results in Table 2).

4.6 Results: In-house data

Having tested our methodology on the UCI OPPORTUNITY dataset, we use our In-house dataset to further assess α-β Network with UQ in a noisier environment. We first measure the activity detection accuracy and F-score of the α-β network, shown in Ta-



(a) Without pre-training (b) With pre-training

Figure 3: Average of α network output in testing without (a) and with (b) pre-training. The model collapses into a single network in the former. This shows the effectiveness of pre-training in making sure that the α network finds subgroups in the data and hence can take advantage of context-aware recognition.

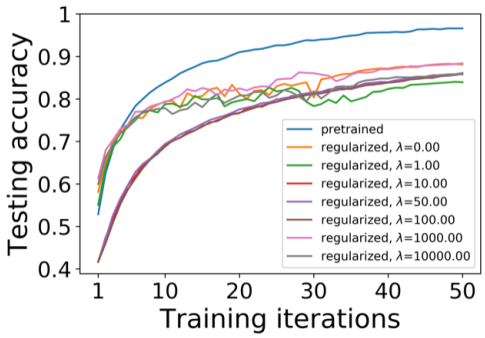


Figure 4: Comparison between pre-training and regularization.

ble 3. It can be seen that our method still outperforms the baseline (84% vs. 80%). Note that our model did drop accuracy compared to the average results from Table 2 (84.3% (std 0.015)), showing we indeed collected a noisier dataset which is more realistic. Secondly, to present the model's ability to detect new contexts on such a dataset, we present the UQ results as well, shown in Figure 5. It is clear that unknown contexts are more spread out, whereas known contexts have more concentrated predictive posterior probability distributions. We have further pushed the experimental setting harder to make it more realistic. That is, in this dataset with three contexts we removed 2 contexts as unknown contexts and just run the UQ solely on the seen contexts, and calculate the probability distribution to see if it can distinguish different contexts. As a result, the model can still find concentrated probability from known contexts but flat distribution for unknown ones, showing our superior performance on unknown context discovery.

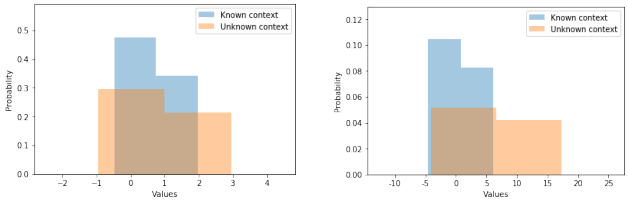


Figure 5: Predicted probability distributions when 1 context is removed vs. 2 contexts are removed

Table 3: In house data, Accuracy and F score for the α-β network and baseline which is a β network with the same number of parameters for each specific number of clusters. For 1 cluster both the models are the same.

Performance Measures	Accuracy	F score
α-β Network	0.84	0.86
baseline	0.80	0.80

5 Conclusion

In this paper, we developed a novel α-β network together with its UQ formulation in tackling a range of realistic situations where context information was unknown but critical for enhanced situation awareness and human activity recognition. Experiments on a real-world dataset showed that this combination of deep learning and uncertainty quantification led to superior performances by its efficacy and efficiency in extracting context information, recognizing unknown contexts, and assessing prediction's uncertainty.

This work can be used as a foundation for a more comprehensive analysis of contextual discovery in a variety of modeling efforts in the future. Multiple levels of contextual information can be gathered and learned from the system, and understanding those in a truly unsupervised setting can enhance a number of recognition tasks and create a flexible and more realistic ontology for how to define context in human activity recognition tasks, rather than relying on high-level but general descriptions of context or too restrictive predefined contexts. In addition, other sources of uncertainty in prediction which result from data of new distributions or noisy data from old distributions should be considered as well.

Acknowledgements

This project is in part supported by the Defense Advanced Research Projects Agency under grand FA8750-18-2-0027 and National Institutes of Health under grant 1R01EB028106-01 and 1R21EB028486-01.

References

Aljundi, R., Chakravarty, P., and Tuytelaars, T. (2017). Expert gate: Lifelong learning with a network of experts. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3366–3375.

Andreu-Perez, J., Leff, D. R., Ip, H. M., and Yang, G.-Z. (2015). From wearable sensors to smart implants—toward pervasive and personalized healthcare. *IEEE Transactions on Biomedical Engineering*, 62(12):2750–2762.

Ardywibowo, R., Zhao, G., Wang, Z., Mortazavi, B., Huang, S., and Qian, X. (2019). Adaptive activity monitoring with uncertainty quantification in switching Gaussian Process models. *The 22nd International Conference on Artificial Intelligence and Statistics (AISTATS)*.

Bauer, D., Kuhnert, L., and Eckstein, L. (2019). Deep, spatially coherent inverse sensor models with uncertainty incorporation using the evidential framework. *arXiv preprint arXiv:1904.00842*.

Bishop, C. M. and Svenskn, M. (2002). Bayesian hierarchical mixtures of experts. In *Proceedings of the Nineteenth conference on Uncertainty in Artificial Intelligence*, pages 57–64. Morgan Kaufmann Publishers Inc.

Cao, L. (2003). Support vector machines experts for time series forecasting. *Neurocomputing*, 51:321–339.

Courbariaux, M., Ambroise, C., Dalmasso, C., Szafranski, M., Consortium, M., et al. (2018). A mixture model with logistic weights for disease subtyping with integrated genome association study.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society. Series B (methodological)*, pages 1–38.

Dey, A. K. (2001). Understanding and using context. *Personal and ubiquitous computing*, 5(1):4–7.

Donahue, J., Jia, Y., Vinyals, O., Hoffman, J., Zhang, N., Tzeng, E., and Darrell, T. (2014). Decaf: A deep convolutional activation feature for generic visual recognition. In *International conference on machine learning*, pages 647–655.

Gal, Y. and Ghahramani, Z. (2016). Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pages 1050–1059.

Guérin, J., Gibaru, O., Thiery, S., and Nyiri, E. (2017). Cnn features are also great at unsupervised classification. *arXiv preprint arXiv:1707.01700*.

Hendrycks, D. and Gimpel, K. (2016). A baseline for detecting misclassified and out-of-distribution examples in neural networks. *arXiv preprint arXiv:1610.02136*.

Jacobs, R. A., Peng, F., and Tanner, M. A. (1997). A bayesian approach to model selection in hierarchical mixtures-of-experts architectures. *Neural Networks*, 10(2):231–241.

Jordan, M. I. and Jacobs, R. A. (1994). Hierarchical mixtures of experts and the em algorithm. *Neural computation*, 6(2):181–214.

Jordan, M. I. and Xu, L. (1995). Convergence results for the em approach to mixtures of experts architectures. *Neural networks*, 8(9):1409–1431.

Lee, K., Lee, K., Lee, H., and Shin, J. (2018). A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In *Advances in Neural Information Processing Systems*, pages 7167–7177.

Liang, S., Li, Y., and Srikant, R. (2017). Principled detection of out-of-distribution examples in neural networks. *arXiv preprint arXiv:1706.02690*.

Lima, C. A., Coelho, A. L., and Von Zuben, F. J. (2007). Hybridizing mixtures of experts with support vector machines: Investigation into nonlinear dynamic systems identification. *Information Sciences*, 177(10):2049–2074.

Lockheed Martin (2019). Anonymized phone usage data collector.

Malinin, A. and Gales, M. (2018). Predictive uncertainty estimation via prior networks. In *Advances in Neural Information Processing Systems*, pages 7047–7058.

Meghdadi, N., Niroomand-Oscuii, H., Soltani, M., Ghalichi, F., and Pourgolmohammad, M. (2017). Brain tumor growth simulation: model validation through uncertainty quantification. *International Journal of System Assurance Engineering and Management*, 8(3):655–662.

Miech, A., Laptev, I., and Sivic, J. (2017). Learnable pooling with context gating for video classification. *arXiv preprint arXiv:1706.06905*.

Nannapaneni, S. and Mahadevan, S. (2014). Uncertainty quantification in performance evaluation of manufacturing processes. In *2014 IEEE International Conference on Big Data (Big Data)*, pages 996–1005. IEEE.

Ravi, N., Dandekar, N., Mysore, P., and Littman, M. L. (2005). Activity recognition from accelerometer data. In *Aaai*, volume 5, pages 1541–1546.

Reynders, E., Maes, K., Lombaert, G., and De Roeck, G. (2016). Uncertainty quantification in operational

modal analysis with stochastic subspace identification: validation and applications. *Mechanical Systems and Signal Processing*, 66:13–30.

Riboni, D. and Bettini, C. (2011). Cosar: hybrid reasoning for context-aware activity recognition. *Personal and Ubiquitous Computing*, 15(3):271–289.

Roggen, D., Calatroni, A., Rossi, M., Holleczek, T., Förster, K., Tröster, G., Lukowicz, P., Bannach, D., Pirkl, G., Ferscha, A., et al. (2010). Collecting complex activity datasets in highly rich networked sensor environments. In *Networked Sensing Systems (INSS), 2010 Seventh International Conference on*, pages 233–240. IEEE.

Samareh, A. and Huang, S. (2019). UQ-CHI: An uncertainty quantification-based contemporaneous health index for degenerative disease monitoring. *arXiv preprint arXiv:1902.08246*.

Sensoy, M., Kaplan, L., and Kandemir, M. (2018). Evidential deep learning to quantify classification uncertainty. In *Advances in Neural Information Processing Systems*, pages 3179–3189.

Sezer, O. B., Dogdu, E., and Ozbayoglu, A. M. (2018). Context-aware computing, learning, and big data in internet of things: a survey. *IEEE Internet of Things Journal*, 5(1):1–27.

Sharif Razavian, A., Azizpour, H., Sullivan, J., and Carlsson, S. (2014). Cnn features off-the-shelf: an astounding baseline for recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 806–813.

Sharma, A., Saxena, S., and Rai, P. (2019). A flexible probabilistic framework for large-margin mixture of experts. *Machine Learning*, pages 1–25.

Shazeer, N., Mirhoseini, A., Maziarz, K., Davis, A., Le, Q., Hinton, G., and Dean, J. (2017). Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *arXiv preprint arXiv:1701.06538*.

Solis, R., Pakbin, A., Akbari, A., Mortazavi, B. J., and Jafari, R. (2019). A human-centered wearable sensing platform with intelligent automated data annotation capabilities. In *Proceedings of the International Conference on Internet of Things Design and Implementation*, pages 255–260. ACM.

Spiegel, B. M., Kaneshiro, M., Russell, M. M., Lin, A., Patel, A., Tashjian, V. C., Zegarski, V., Singh, D., Cohen, S. E., Reid, M. W., et al. (2014). Validation of an acoustic gastrointestinal surveillance biosensor for postoperative ileus. *Journal of Gastrointestinal Surgery*, 18(10):1795–1803.

Steven Eyobu, O. and Han, D. (2018). Feature representation and data augmentation for human activity classification based on wearable imu sensor data using a deep lstm neural network. *Sensors*, 18(9):2892.

Sun, S., Liu, Y., and Mao, L. (2018). Multi-view learning for visual violence recognition with maximum entropy discrimination and deep features. *Information Fusion*.

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

Xu, J., Song, L., Xu, J. Y., Pottie, G. J., and Van Der Schaar, M. (2016). Personalized active learning for activity classification using wireless wearable sensors. *IEEE journal of selected topics in signal processing*, 10(5):865–876.

Xu, J. Y., Chang, H.-I., Chien, C., Kaiser, W. J., and Pottie, G. J. (2014). Context-driven, prescription-based personal activity classification: methodology, architecture, and end-to-end implementation. *IEEE journal of biomedical and health informatics*, 18(3):1015–1025.

Yuan, C. and Neubauer, C. (2009). Variational mixture of gaussian process experts. In *Advances in Neural Information Processing Systems*, pages 1897–1904.

Yuksel, S. E., Wilson, J. N., and Gader, P. D. (2012). Twenty years of mixture of experts. *IEEE transactions on neural networks and learning systems*, 23(8):1177–1193.

Zhu, C. and Wang, Z. (2018). Semi-supervised soft margin consistency based multi-view maximum entropy discrimination. *Applied Computing and Informatics*.