VFDS: Variational Foresight Dynamic Selection in Bayesian Neural Networks for Efficient Human Activity Recognition

Randy Ardywibowo[1] **Shahin Boluki**[1] **Zhangyang Wang**[2] **Bobak Mortazavi**[1]
Shuai Huang[3] **Xiaoning Qian**[1]
Texas A&M University[1] University of Texas at Austin[2] University of Washington[3]

Abstract

In many machine learning tasks, input features with varying degrees of predictive capability are acquired at varying costs. In order to optimize the performance-cost trade-off, one would select features to observe *a priori*. However, given the changing context with previous observations, the subset of predictive features to select may change dynamically. Therefore, we face the challenging new problem of *foresight dynamic selection* (FDS): finding a dynamic and light-weight policy to decide which features to observe next, **before actually observing them**, for overall performance-cost trade-offs. To tackle FDS, this paper proposes a Bayesian learning framework of *Variational Foresight Dynamic Selection* (**VFDS**). VFDS learns a policy that selects the next feature subset to observe, by optimizing a variational Bayesian objective that characterizes the trade-off between model performance and feature cost. At its core is an implicit variational distribution on binary gates that are dependent on previous observations, which will select the next subset of features to observe. We apply VFDS on the Human Activity Recognition (HAR) task where the performance-cost trade-off is critical in its practice. Extensive results demonstrate that VFDS selects different features under changing contexts, notably saving sensory costs while maintaining or improving the HAR accuracy. Moreover, the features that VFDS dynamically select are shown to be interpretable and associated with the different activity types.

Proceedings of the 25[th] International Conference on Artificial Intelligence and Statistics (AISTATS) 2022, Valencia, Spain. PMLR: Volume 151.

1 INTRODUCTION

Acquiring predictive features is critical for building trustworthy machine learning systems, but this may come at a daunting cost. Such a cost can be in the form of energy needed to maintain an ambient sensor (Ardywibowo et al., 2019, 2018; Yang et al., 2020), time needed to complete an experiment (Kiefer, 1959), or manpower required to monitor a hospital patient (Pierskalla and Brailer, 1994; Jiang et al., 2019). It is important not only to maintain good performance in the specified task, but also a low cost to gather features.

For example, existing Human Activity Recognition (HAR) methods typically use a fixed set of sensors, potentially collecting redundant features to discriminate contexts and/or activity types (Shen and Varshney, 2013; Aziz, Robinovitch, and Park, 2016; Ertuğrul and Kaya, 2017; Cheng et al., 2018; Ardywibowo, 2017). Classic feature selection methods such as the LASSO and its variants can address the performance-cost trade-off by optimizing an objective penalized by a term that helps promote feature sparsity (Tibshirani, 1996; Friedman, Hastie, and Tibshirani, 2010, 2008; Zou and Hastie, 2005). Such feature selection formulations are often static, i.e., a fixed set of features are selected *a priori*. However, different features may offer different predictive power under different contexts. For example, a health worker may not need to monitor a recovering patient as frequently as a patient with declining conditions; or a smartphone sensor may be predictive when the user is walking but not in a car. By dynamically selecting which feature(s) to observe, one can further reduce the inherent cost for prediction and achieve a better trade-off between cost and prediction accuracy.

In addition to cost-efficiency, a dynamic feature selection formulation can also lead to more interpretable and trustworthy predictions. Specifically, the predictions made by the model are only based on the selected features, providing a clear relationship between input

features and model predictions. Existing efforts on interpreting models are usually based on some post-analyses of the predictions, including the approaches in (1) visualizing higher-level representations or reconstructions of inputs based on them (Li et al., 2016b; Mahendran and Vedaldi, 2015), (2) evaluating the sensitivity of predictions to local perturbations of inputs or input gradients (Selvaraju et al., 2017; Ribeiro, Singh, and Guestrin, 2016), and (3) extracting parts of inputs as justifications for predictions (Lei, Barzilay, and Jaakkola, 2016). Another related but orthogonal direction is model compression: training sparse neural networks with the goal of memory and computational efficiency (Louizos, Welling, and Kingma, 2017; Tartaglione et al., 2018; Han et al., 2015). All these works require collecting all features first and provide post-hoc feature or model pruning.

Recent efforts on dynamic feature selection select which features to observe based on immediate statistics (Gordon et al., 2012; Bloom, Argyriou, and Makris, 2013; Ardywibowo et al., 2019; Zappi et al., 2008), ignoring the information a feature may have on future predictions. Others treat feature selection as a Markov Decision Process (MDP) and use Reinforcement Learning (RL) to solve it (He and Eisner, 2012; Karayev, Fritz, and Darrell, 2013; Kolamunna et al., 2016; Spaan and Lima, 2009; Satsangi, Whiteson, and Oliehoek, 2015; Yang et al., 2020). However, solving RL is not straightforward. Besides being sensitive to hyperparameter settings in general, approximations such as state space discretization and relaxation of the combinatorial objective were used to make the RL problem tractable.

On the other hand, Bayesian inference offers a way to learn a model that formalizes our dynamic feature selection hypothesis. In this direction, Koop and Korobilis (2018) proposed using simple variational distributions to dynamically select predictive models; however, the method is limited to linear, time-varying parameter models. Meanwhile, Bayesian Neural Networks (BNNs) offer a method of training Neural Networks (NNs) while preventing them from overfitting. These methods treat the NN weights as random variables and regularize them with appropriate prior distributions (MacKay, 1992; Neal, 2012). To scale these techniques to real-world applications, various types of approximate inference techniques have been developed (Graves, 2011; Welling and Teh, 2011; Li et al., 2016a; Blundell et al., 2015; Louizos and Welling, 2017; Shi, Sun, and Zhu, 2018; Gal and Ghahramani, 2016; Gal, Hron, and Kendall, 2017). In particular, (semi-)implicit variational inference offers a way to define expressive distributions to better approximate the posterior distribution, enabling more complex models to be inferred efficiently (Yin and Zhou, 2018b; Titsias and Ruiz, 2019; Molchanov et al., 2019). Despite this, the extension of these methods for dynamic feature selection in BNNs has not been explored before. We refer the reader to the supplementary materials for additional related work on static selection, dynamic selection, and variational inference.

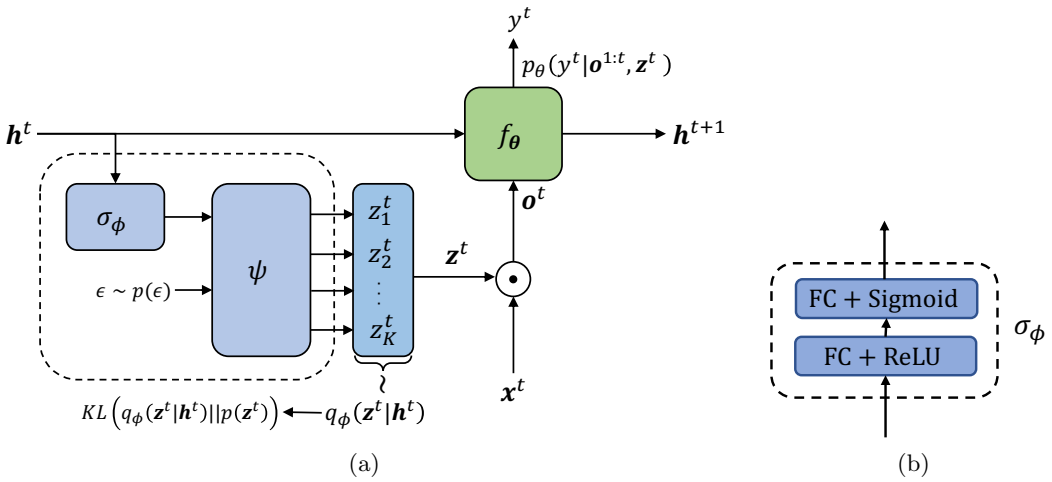
To this end, we propose VFDS, a variational dynamic feature selection method for Bayesian Neural Networks that can be easily used with existing deep architecture components and trained from end-to-end, enabling *task-driven dynamic feature selection*. To achieve this, we first define a prior distribution on binary random variables that determines which features to observe next in order to characterize the model performance-cost trade-off. We then design an implicit variational distribution on the binary variables that is conditioned on previous observations, allowing us to dynamically select features at any given time based on previous observations. Through stochastic approximations and differentiable relaxations, we are able to jointly optimize the parameters of this distribution along with the model parameters with respect to the variational objective. To show our method's ability to dynamically select features while maintaining good performance, we evaluate it on four time-series activity recognition datasets: the UCI Human Activity Recognition (HAR) dataset (Anguita et al., 2013), the OPPORTUNITY dataset (Roggen et al., 2010), the ExtraSensory dataset (Vaizman, Ellis, and Lanckriet, 2017), as well as the NTU-RGB-D dataset (Shahroudy et al., 2016).

Several ablation studies and comparisons with other dynamic and static feature selection methods demonstrate the efficacy of our proposed method. In some cases, VFDS only needs to observe 0.28% features on average while still maintaining competitive human activity monitoring accuracy, compared to 14.38% used by static feature selection methods. This indicates that some features are indeed redundant for certain contexts. We further show that the dynamically selected features are shown to be interpretable with direct correspondence with different contexts and activity types.

2 METHODOLOGY

We define our notation as follows: let $\mathcal{D}$ be a dataset containing N independent and identically distributed (*iid*) input-output pairs of *multivariate* time series data $\{(\boldsymbol{x}_1, \boldsymbol{y}_1), \ldots, (\boldsymbol{x}_N, \boldsymbol{y}_N)\}$ of length $\{T_1, \ldots, T_N\}$. For each time point t, $\boldsymbol{x}_i^t$ is a data point containing K features $\{\boldsymbol{x}_{i,1}^t, \ldots, \boldsymbol{x}_{i,K}^t\}$, and y_i^t is a target output for that time point.

We are interested in learning a model that uses the previously observed data to infer the feature set we should observe next, as well as predicting the target



(a) (b)

Figure 1: **(a)** A variational foresight dynamic feature selection (VFDS) module illustrated for one timestep. $\boldsymbol{h}^t$ is used to determine the variational distribution $q_{\boldsymbol{\phi}}(\boldsymbol{z}^t|\boldsymbol{h}^t)$ from which the feature selection gates $\boldsymbol{z}^t$ are drawn. $q_{\boldsymbol{\phi}}(\boldsymbol{z}^t|\boldsymbol{h}^t)$ is defined implicitly through a transformation of a random variable $\boldsymbol{\epsilon}$ drawn from an explicit distribution $p(\boldsymbol{\epsilon})$, by the covariate-dependent function $\boldsymbol{\psi}_{\boldsymbol{\phi}}(\boldsymbol{h}^t, \boldsymbol{\epsilon})$. By choosing $\boldsymbol{\psi}_{\boldsymbol{\phi}}(\cdot)$ carefully, we can have $\mathbb{E}_{p(\boldsymbol{\epsilon})}[\boldsymbol{\psi}_{\boldsymbol{\phi}}(\boldsymbol{h}^t, \boldsymbol{\epsilon})] = \boldsymbol{\sigma}_{\boldsymbol{\phi}}(\boldsymbol{h}^t)$, where $\boldsymbol{\sigma}_{\boldsymbol{\phi}}(\cdot)$ can be defined by a neural network. We can obtain a closed form approximation of the KL divergence. The gates are then used to determine which features to observe for the current time-step. **(b)** The neural network architecture used for $\boldsymbol{\sigma}_{\boldsymbol{\phi}}(\boldsymbol{h})$.

output for the next time point. We hypothesize that time-series predictive models for human activity recognition do not require all features be observed at all times. Indeed, many features in multivariate sensor data may be redundant for prediction in a given context. Thus, dynamically choosing which features to observe at any given time would be beneficial as sensors can be dynamically turned on or off depending on specific monitoring needs. Moreover, dynamic feature selection may enable better interpretability on which sensors are required for any given context.

We formalize this hypothesis under the Bayesian Neural Network (BNN) learning paradigm. Deep neural networks offer high predictive capability on complex datasets, allowing us to achieve high performance on the monitoring task. On the other hand, Bayesian statistics offer a way to formalize our hypothesis and learn these neural networks without overfitting. In subsequent sections, we formulate a Bayesian learning problem for foresight dynamic feature selection in terms of variational inference.

2.1 Variational Objective of VFDS

Bayesian learning can be formulated as maximizing a log-marginal likelihood: $\log p(\boldsymbol{y}|\boldsymbol{x}) = \log \prod_{i=1}^N p(\boldsymbol{y}_i|\boldsymbol{x}_i) = \log \int \prod_{i=1}^N p(\boldsymbol{y}_i|\boldsymbol{x}_i, \boldsymbol{z}) p(\boldsymbol{z}) d\boldsymbol{z}$, where $\boldsymbol{z}$ are the intermediary parameters of the model, considered as random variables. This log-marginal is often intractable, and it is common to resort to variational inference by introducing a variational distribution $q(\boldsymbol{z})$ on the parameters of the model. With this, maximizing the log-marginal is often transformed to minimizing the negative Evidence Lower Bound (ELBO) (Hoffman et al., 2013; Blei, Kucukelbir, and McAuliffe, 2017). In the case of time-series data with partial observations, the negative ELBO can be written as follows:

$$\mathcal{L}(\mathcal{D}) = -\sum_{i=1}^{N}\sum_{t=1}^{T_i} \mathbb{E}_{\boldsymbol{z}\sim q(\boldsymbol{z}), \boldsymbol{\theta}\sim q(\boldsymbol{\theta})}[\log p(y_i^t|\boldsymbol{o}_i^{1:t}, \boldsymbol{z}, \boldsymbol{\theta})] + \mathrm{KL}(q(\boldsymbol{z})||p(\boldsymbol{z})), \tag{1}$$

where $\boldsymbol{o}_i^{1:t}$ are the observed features up to time t. Under this framework, we can form a variational approximation to training dynamic feature selection with deep networks. We do this by introducing stochastic, input-dependent binary variables $z_{i,k}^t$ that determine whether feature k is observed at time t.

To simplify our exposition, we focus on selecting features for one instance i at time-point t, and omit these subscripts in our exposition, reintroducing them later for clarity. We present the inference of our binary selection variables. A fully Bayesian treatment of the other neural network model parameters $\boldsymbol{\theta}$ can be considered through standard approximate Bayesian inference techniques for neural networks such as Monte Carlo (MC) dropout and its variants (Gal and Ghahramani, 2016; Gal, Hron, and Kendall, 2017; Kingma, Salimans, and Welling, 2015; Boluki et al., 2020). For each feature k, we define a prior distribution $p(z_k)$ for each binary variable as

$$p(z_k) = \texttt{Bern}(e^{-\eta c_k}). \tag{2a}$$

Here, c_k is the energy cost of feature k, and η is a

parameter controlling the shape of the prior. We then define an implicit, covariate-dependent variational distribution that is dependent on a belief state $\boldsymbol{h}$ at time t that summarizes the previous observations. Specifically, the variational distribution $q(\boldsymbol{z}|\boldsymbol{h}) = q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{h})$ with parameters $\boldsymbol{\phi}$ is defined by transforming random variables from an explicit distribution $\boldsymbol{\epsilon} \sim p(\boldsymbol{\epsilon})$ using a reparameterizable transformation as follows (Kingma and Welling, 2013; Titsias and Ruiz, 2019):

$$\boldsymbol{\epsilon} \sim p(\boldsymbol{\epsilon}), \quad \boldsymbol{z} = \boldsymbol{\psi}_{\boldsymbol{\phi}}(\boldsymbol{h}, \boldsymbol{\epsilon}) \quad \equiv \quad \boldsymbol{z} \sim q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{h}). \tag{3}$$

Here, $\boldsymbol{\psi}_{\boldsymbol{\phi}}(\cdot)$ outputs a binary random vector that determines whether feature k is selected, where $z_k = \boldsymbol{\psi}_{\boldsymbol{\phi}}(\boldsymbol{h}, \boldsymbol{\epsilon})_k$. The details of this transformation will be explained in the following sections.

With this, the first term of $\mathcal{L}(\mathcal{D})$ can be estimated by using a single sample of $\boldsymbol{z}$ for each time point t of instance i. On the other hand, the KL term can be computed as

$$\mathrm{KL}(q(\boldsymbol{z}|\boldsymbol{h})||p(\boldsymbol{z})) = \sum_{k=1}^{K} \mathrm{KL}(q_{\boldsymbol{\phi}}(z_k|\boldsymbol{h})||p(z_k)), \tag{4}$$

$$\begin{aligned}\mathrm{KL}(q_{\boldsymbol{\phi}}(z_k|\boldsymbol{h})||p(z_k)) = -H[q_{\boldsymbol{\phi}}(z_k|\boldsymbol{h})] + \\ \eta c_k q_{\boldsymbol{\phi}}(z_k = 1|\boldsymbol{h}) - \log\left(1 - e^{-\eta c_k}\right) q_{\boldsymbol{\phi}}(z_k = 0|\boldsymbol{h}),\end{aligned} \tag{5}$$

where $H[q_{\boldsymbol{\phi}}(z_k|\boldsymbol{h})]$ is the entropy of $q_{\boldsymbol{\phi}}(z_k|\boldsymbol{h})$. For sufficiently large η, $\log\left(1 - e^{-\eta c_k}\right) \approx 0$. We achieve this by scaling η with N, $\eta = N\lambda$. We can then scale the negative ELBO with the number of samples N without changing the optima:

$$\begin{aligned}\mathrm{KL}(q_{\boldsymbol{\phi}}(z_k|\boldsymbol{h})||p(z_k)) \approx \\ -\frac{1}{N} H[q_{\boldsymbol{\phi}}(z_k|\boldsymbol{h})] + \frac{N\lambda}{N} c_k q_{\boldsymbol{\phi}}(z_k = 1|\boldsymbol{h}).\end{aligned} \tag{6}$$

For large N, the entropy term vanishes, leaving us with

$$\mathrm{KL}(q_{\boldsymbol{\phi}}(z_k|\boldsymbol{h})||p(z_k)) \approx \lambda c_k q_{\boldsymbol{\phi}}(z_k = 1|\boldsymbol{h}). \tag{7}$$

Note that $q_{\boldsymbol{\phi}}(z_k = 1|\boldsymbol{h}) = \mathbb{E}_{p(\boldsymbol{\epsilon})}[\boldsymbol{\psi}_{\boldsymbol{\phi}}(\boldsymbol{h}, \boldsymbol{\epsilon})_k]$. By defining $p(\boldsymbol{\epsilon})$ as the logistic distribution with probability density $f(\boldsymbol{\epsilon})$, and $\boldsymbol{\psi}_{\boldsymbol{\phi}}(\boldsymbol{h}, \boldsymbol{\epsilon})$ through a deterministic function $\boldsymbol{\sigma}_{\boldsymbol{\phi}}(\boldsymbol{h}) \in (0, 1)$ as follows:

$$\boldsymbol{\epsilon} \sim p(\boldsymbol{\epsilon}), \quad f(\boldsymbol{\epsilon}) = \frac{e^{-\boldsymbol{\epsilon}}}{(1 + e^{-\boldsymbol{\epsilon}})^2}, \tag{8}$$

$$\boldsymbol{\psi}_{\boldsymbol{\phi}}(\boldsymbol{h}, \boldsymbol{\epsilon}) = \mathbb{1}\left[\log\left(\frac{\boldsymbol{\sigma}_{\boldsymbol{\phi}}(\boldsymbol{h})}{1 - \boldsymbol{\sigma}_{\boldsymbol{\phi}}(\boldsymbol{h})}\right) + \boldsymbol{\epsilon} > 0\right], \tag{9}$$

we have that $\mathbb{E}_{p(\boldsymbol{\epsilon})}[\boldsymbol{\psi}_{\boldsymbol{\phi}}(\boldsymbol{h}, \boldsymbol{\epsilon})] = \boldsymbol{\sigma}_{\boldsymbol{\phi}}(\boldsymbol{h})$. In practice, $\boldsymbol{\epsilon} \sim p(\boldsymbol{\epsilon})$ can be sampled as $\boldsymbol{\epsilon} = \log \boldsymbol{u} - \log(1 - \boldsymbol{u})$, where $\boldsymbol{u} \sim \texttt{Unif}(0, 1)$. On the other hand, we specify $\boldsymbol{\sigma}_{\boldsymbol{\phi}}(\boldsymbol{h})$ as a neural network, whose architecture we will describe in later sections. Now, our approximation to the KL term can be written as

$$\mathrm{KL}(q_{\boldsymbol{\phi}}(z_k|\boldsymbol{h})||p(z_k)) \approx \lambda c_k \boldsymbol{\sigma}_{\boldsymbol{\phi}}(\boldsymbol{h})_k. \tag{10}$$

By using this approximation in the negative ELBO $\mathcal{L}(\mathcal{D})$, rewriting the ELBO in terms of an expectation, and reintroducing subscripts, we arrive at the following objective:

$$\begin{aligned}\mathcal{L}(\mathcal{D}) = \mathbb{E}_{(\boldsymbol{x}, \boldsymbol{y}, i) \sim \mathcal{D}}\Bigg[\\ - \sum_{t=1}^{T_i} \mathbb{E}_{\boldsymbol{z}_i^t \sim q_{\boldsymbol{\phi}}(\boldsymbol{z}_i^t|\boldsymbol{h}_i^t), \boldsymbol{\theta} \sim q_{\boldsymbol{\phi}}(\boldsymbol{\theta})}\left[\log p(y^t|\boldsymbol{x}^t, \boldsymbol{z}_i^t, \boldsymbol{\theta})\right] \\ + \lambda \sum_{t=1}^{T_i} \sum_{k=1}^{K} c_k \boldsymbol{\sigma}_{\boldsymbol{\phi}}(\boldsymbol{h}_i^t)_k \Bigg].\end{aligned} \tag{11}$$

We see here that there are two terms in our objective function, the first term is a likelihood term that determines how accurately the model recovers the target distribution, and the second term penalizes the model for dynamically choosing to observe too many features on average at each time-point, weighted by their energy cost. Intuitively, $\boldsymbol{\sigma}_{\boldsymbol{\phi}}(\boldsymbol{h})$ can be thought of as a gating module that selects features to observe based on the previous observations. In the following sections, we describe the architecture of this gating mechanism in greater details, as well as practical considerations when attempting to optimize with respect to this objective and apply this method in practice.

2.2 Foresight Dynamic Selection Module

We now describe our dynamic feature selection module, which can be seen in Figure 1. Here, we adopt a Recurrent Neural Network (RNN) structure, using $f_{\boldsymbol{\theta}}(\cdot)$ with parameters $\boldsymbol{\theta}$ to compute the belief state $\boldsymbol{h}^t$ for a given time t (Graves, Mohamed, and Hinton, 2013). We then use the hidden state $\boldsymbol{h}$ to implicitly define the distribution $q_{\boldsymbol{\phi}}(\boldsymbol{z}^t|\boldsymbol{h}^t)$ from which we sample the feature selection gates $\boldsymbol{z}^t$. This is done by first feeding $\boldsymbol{h}^t$ through a gating module $\sigma_{\boldsymbol{\phi}}(\boldsymbol{h}^t)$ with variational parameters $\boldsymbol{\phi}$. This module is defined by a neural network consisting of two fully connected layers with ReLU and sigmoid activation functions, respectively. This can be seen in Figure 1(b). We then use the output of this module to transform the random variable ϵ into $\boldsymbol{z}^t$, thereby sampling $\boldsymbol{z}^t$ from $q_{\boldsymbol{\phi}}(\boldsymbol{z}^t|\boldsymbol{h}^t)$.

With this, our optimization problem aims at minimizing $\mathcal{L}(\mathcal{D})$ by optimizing the parameters $\boldsymbol{\theta}$ and variational parameters $\boldsymbol{\phi}$. We intend to solve this problem through gradient-based methods. However, the discrete random variables $\boldsymbol{z}^t$'s are not directly amenable to stochastic

reparameterization techniques. In the following, we describe a differentiable relaxation that we adopt to allow the training of our method end-to-end, enabling easy integration into many existing deep architectures.

2.3 Differentiable Relaxation

The final hurdle in solving the above problem using gradient descent is that the discrete random variables $\boldsymbol{z}^t$'s are not directly amenable to stochastic reparameterization techniques. An effective and simple to implement formulation that we adopt is the Gumbel-Softmax reparameterization (Jang, Gu, and Poole, 2016; Maddison, Mnih, and Teh, 2016; Ardywibowo et al., 2020). It relaxes a discrete valued random variable $\boldsymbol{z}$ to a continuous random variable $\tilde{\boldsymbol{z}}$. Specifically, the discrete valued random variables $\boldsymbol{z}$ can instead be relaxed into continuous random variables $\tilde{\boldsymbol{z}}$ through the transformation $\tilde{\boldsymbol{\psi}}_{\boldsymbol{\phi}}(\boldsymbol{x}, \boldsymbol{\epsilon})$ as follows:

$$\tilde{\boldsymbol{\psi}}_{\boldsymbol{\phi}}(\boldsymbol{x}, \boldsymbol{\epsilon}) = \textsc{Sigmoid}\left(\left(\log\left(\frac{\boldsymbol{\sigma}_{\boldsymbol{\phi}}(\boldsymbol{x})}{1-\boldsymbol{\sigma}_{\boldsymbol{\phi}}(\boldsymbol{x})}\right) + \boldsymbol{\epsilon}\right) \Big/ \tau\right), \tag{12}$$

where $\boldsymbol{\epsilon}$ is a sample from a logistic distribution defined in Section 2.1. Meanwhile, τ is a temperature parameter. For low values of τ, $\boldsymbol{s}$ approaches a sample of a binary random variable, recovering the original discrete problem, while for high values, $\boldsymbol{s}$ will equal $\frac{1}{2}$.

With this, we are able to compute gradient estimates of $\tilde{\boldsymbol{z}}$ and approximate the gradient of $\boldsymbol{z}$ as $\nabla_{\boldsymbol{\theta},\boldsymbol{\phi}}\boldsymbol{z} \approx \nabla_{\boldsymbol{\theta},\boldsymbol{\phi}}\tilde{\boldsymbol{z}}$. This enables us to backpropagate through the discrete random variables and train the selection parameters along with the model parameters jointly using stochastic gradient descent. At test time, we remove the stochasticity and set the gates as $\boldsymbol{z} = \mathbb{1}\big[\log\big(\frac{\boldsymbol{\sigma}_{\boldsymbol{\phi}}(\boldsymbol{x})}{1-\boldsymbol{\sigma}_{\boldsymbol{\phi}}(\boldsymbol{x})}\big) > 0\big]$, or equivalently, set $\boldsymbol{z} = \mathbb{1}[\boldsymbol{\sigma}_{\boldsymbol{\phi}}(\boldsymbol{x}) > \frac{1}{2}]$.

We can see that such a module can be easily integrated into many existing deep architectures and trained from end-to-end, enabling *task-driven feature selection*. We demonstrate this ability by applying it to a variety of recurrent architectures such as a Gated Recurrent Unit (GRU) (Cho et al., 2014) and an Independent RNN (Li et al., 2018).

3 EXPERIMENTS

We evaluate VFDS on four different datasets: the UCI Human Activity Recognition (HAR) using Smartphones Dataset (Anguita et al., 2013), the OPPORTUNITY Dataset (Roggen et al., 2010), the ExtraSensory dataset (Vaizman, Ellis, and Lanckriet, 2017), and the NTU-RGB-D dataset (Shahroudy et al., 2016). Although there are many other human activity recognition benchmark datasets (Chen et al., 2020), we choose the above datasets to better convey our message of achieving feature usage efficiency and interpretability using our dynamic feature selection framework with the following reasons. First, the UCI HAR dataset is a clean dataset with no missing values, allowing us to benchmark different methods without any discrepancies in data preprocessing confounding our evaluations. Second, the OPPORTUNITY dataset contains activity labels that correspond to specific sensors. An optimal dynamic feature selector should primarily choose these sensors under specific contexts with clear physical meaning. The ExtraSensory dataset studies a multilabel classification problem, where two or more labels can be active at any given time. Finally, the NTU-RGB-D dataset is a large-scale activity recognition dataset with over 60 classes of activities using data from 25 skeleton joints, allowing us to benchmark model performance in a complex setting. For all datasets, we randomly split data both chronologically and by different subjects.

We investigate several aspects of our model performance on these benchmarks. To show the effect in prediction accuracy when our selection module is considered, we compare its performance to a standard GRU architecture (Cho et al., 2014). To show the effect of considering dynamic feature selection, we compare a static feature selection formulation using the technique by Louizos, Welling, and Kingma (2017). To benchmark the performance of our differentiable relaxation-based optimization strategy, we implement the Straight-Through estimator (Hinton, Srivastava, and Swersky, 2012), ℓ_1 relaxed regularization, and Augment-REINFORCE-Merge (ARM) gradient estimates (Yin and Zhou, 2018a) as alternative methods to optimize our formulation. The fully sequential application of ARM was not addressed in the original paper, and will be prohibitively expensive to compute exactly. Hence, we combine ARM and Straight-Through (ST) estimator (Hinton, Srivastava, and Swersky, 2012) as another approach to optimize our formulation. More specifically, we calculate the gradients with respect to the Bernoulli variables with ARM, and use the ST

Table 1: Comparison of various optimization techniques for our model on the UCI HAR dataset. *Accuracy and average number of features selected are in (%).

Method	Accuracy	Feat. Selected
ℓ_1 Regularization	90.43	19.48
Straight Through	89.38	0.31
ARM	95.73	11.67
ST-ARM	92.79	1.92
Gumbel-Softmax	**97.18**	**0.28**

Table 2: Comparison of various models for dynamic feature selection on three activity recognition datasets. *Accuracy metrics and average number of features selected are all in (%).

Method	UCI HAR		OPPORTUNITY		ExtraSensory		
	Accuracy	Features	Accuracy	Features	Accuracy	F1	Features
No Selection (GRU)	96.67	100	84.16	100	**91.14**	53.53	100
Static	95.49	14.35	81.63	49.57	91.13	53.18	42.32
Random	52.46	25.00	34.11	50.00	39.66	23.53	40.00
MDP	62.21	24.58	44.45	34.68	48.20	28.11	31.98
IDSS	87.88	10.39	72.95	28.50	70.59	33.24	22.07
Attention	**98.38**	49.94	83.42	54.20	90.37	53.29	54.73
VFDS	97.18	**0.28**	**84.26**	**15.88**	**91.14**	**55.06**	**11.25**

estimator to backpropagate the gradients through the Bernoulli variables to previous layers' parameters. We further compare with an attention-based feature selection, selecting features based on the largest attention weights. Because attention yields feature attention weights instead of feature subsets, we select features by using a hard threshold α of the attention weights and scaling the selected features by $1-\alpha$ for different values of α. Indeed, without this modification, we observe that an attention-based feature selection would select 100% of the features at all times. As additional benchmarks, we compare against a random selection baseline, a Markov Decision Process (MDP), and IDSS, a Reinforcement Learning (RL) based method by Yang et al. (2020).

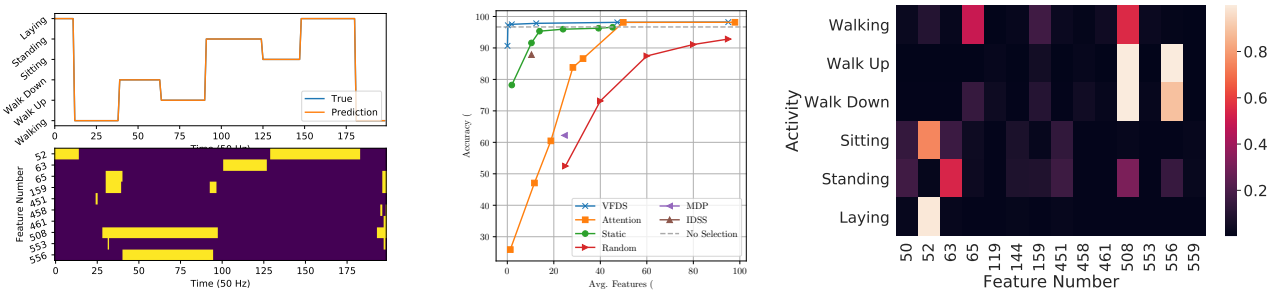
We also have tested different values for the temperature hyperparameter τ, where we observe that the settings with τ below 1 generally yield the best results with no noticeable performance difference. This experiment, discussions on the ExtraSensory dataset, and additional experiments on the stability of our dynamic selection formulation can be found in the supplementary materials.

UCI HAR Dataset: We test our proposed method on performing simultaneous prediction and dynamic feature selection on the UCI HAR dataset (Anguita et al., 2013). This dataset consists of 561 smartphone sensor measurements including various gyroscope and accelerometer readings, with the task of inferring the activity that the user performs at any given time. There are six possible activities that a subject can perform: walking, walking upstairs, walking downstairs, sitting, standing, and laying. Additional experiment details can be found in the supplementary materials.

We first compare various optimization methods using stochastic gradients by differential relaxation via Gumbel-Softmax (GS) reparametrization, Straight-Through (ST), ARM, ST-ARM gradients, and an ℓ_1 regularized formulation to solve dynamic feature selection. As shown by the results provided in Table 1, Gumbel-Softmax achieves the best prediction accuracy with the least number of features. Utilizing either ST, ARM, or ST-ARM for gradient estimation cannot provide a better balance between accuracy and efficiency compared with the Gumbel-Softmax relaxation-based optimization. Indeed, the performance of the ST estimator is expected, as there is a mismatch between the forward propagated activations and the backward propagated gradients in the estimator. Meanwhile, we attribute the worse performance of the ARM and ST-ARM optimizer to its use in a sequential fashion, which was not originally considered. The lower performance of the ℓ_1 regularized formulation is expected as ℓ_1 regularization is an approximation to the optimal feature subset selection problem.

Benchmarking results of different models are given in Table 2. As shown, our dynamic feature selection model is able to achieve a competitive accuracy using only 0.28% of the features, or on average about 1.57 sensors at any given time. We also observe that both the attention and our dynamic formulation are able to improve upon the accuracy of the standard GRU, suggesting that feature selection can also regularize the model to improve accuracy. Although the attention-based model yields the best accuracy, about 50% of the features are used on average compared to 0.28% for our method.

We study the effect of the regularization weight λ by varying it from $\lambda \in \{1, 0.1, 0.01, 0.005, 0.001\}$. We compare this with the attention model by varying the threshold α used to select features from $\alpha \in \{0.5, 0.9, 0.95, 0.99, 0.995, 0.999\}$, as well as the static selection model by varying its λ from $\lambda \in \{1, 0.1, \ldots 0.01, 0.005, 0.001\}$. A trade-off curve between the number of selected features and the performance for the three models can be seen in Figure 2(b). As shown in the figure, the accuracy of the attention model suffers increasingly with smaller feature subsets, as attention is not a formulation specifically tailored to find sparse solutions. On the other hand, the accuracy of our dynamic formulation is unaffected by the number



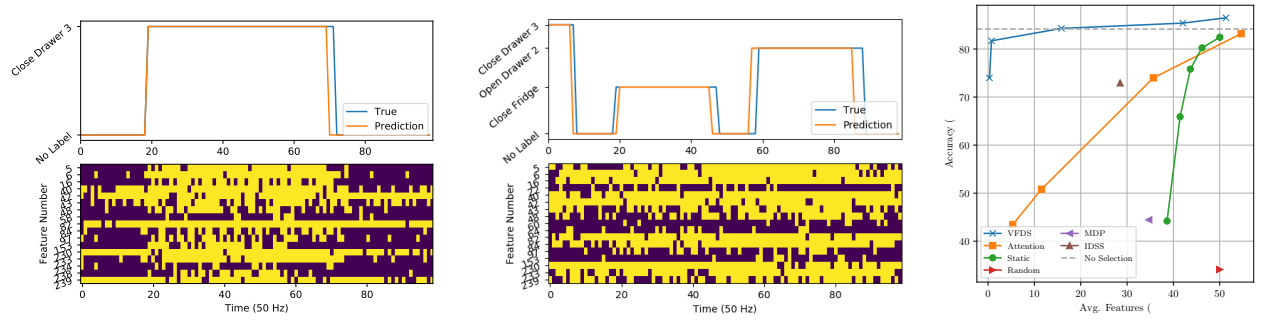
(a) (b) (c)

Figure 2: UCI HAR Dataset results: **(a)** Prediction and features selected of the proposed model $\lambda = 1$. **(b)** Feature selection vs. accuracy trade-off curve comparison. **(c)** Heatmap of sensor feature activations under each activity of the UCI HAR dataset. Only active features are shown out of the 561 features in total.

of features, suggesting that selecting around 0.3% of the features on average may be optimal for the given problem. It further confirms that our dynamic formulation selects the most informative features given the context. Moreover, as we show in the supplementary materials, some features are not selected at all by our dynamic feature selector. The performance of the static selection model is consistent for feature subsets of size 10% or greater. However, it suffers a drop in accuracy for extremely small feature subsets. This shows that for static selection, selecting too many features may result in collecting redundant ones for certain contexts, while selecting a feature set that is too small would be insufficient for maintaining accuracy.

An example of dynamically selected features can be seen in Figure 2(a). We plot the prediction of our model compared to the true label and illustrate the features that are used for prediction. We also plot a heatmap for the features selected under each activity in Figure 2(c). Although these features alone may not be exclusively attributed as the only features necessary for prediction under specific activities, such a visualization is useful to retrospectively observe the features selected by our model at each time-point. Note that mainly 5 out of the 561 features are used for prediction at any given time. Observing the selected features, we see that for the static activities such as sitting, standing, and laying, only sensor feature 52 and 63, features relating to the gravity accelerometer, are necessary for prediction. On the other hand, the active states such as walking, walking up, and walking down requires 3 sensor features: sensor 65, 508, and 556, which are related to both the gravity accelerometer and the body accelerometer. This is intuitively appealing as, under the static contexts, the body accelerometer measurements would be relatively constant, and unnecessary for prediction. On the other hand, for the active contexts, the body accelerometer measurements are necessary to reason about how the subject is moving and accurately discriminate between the different active states. Meanwhile, we found that measurements relating to the gyroscope were unnecessary for prediction.

OPPORTUNITY Dataset: We further test our proposed method on the UCI OPPORTUNITY Dataset



(a) (b) (c)

Figure 3: OPPORTUNITY Dataset results: **(a)** Prediction and features selected of the proposed model $\lambda = 1$. **(b)** Prediction and features selected of the proposed model on a set of activity transitions. **(c)** Feature selection vs. Error trade-off curve comparison.

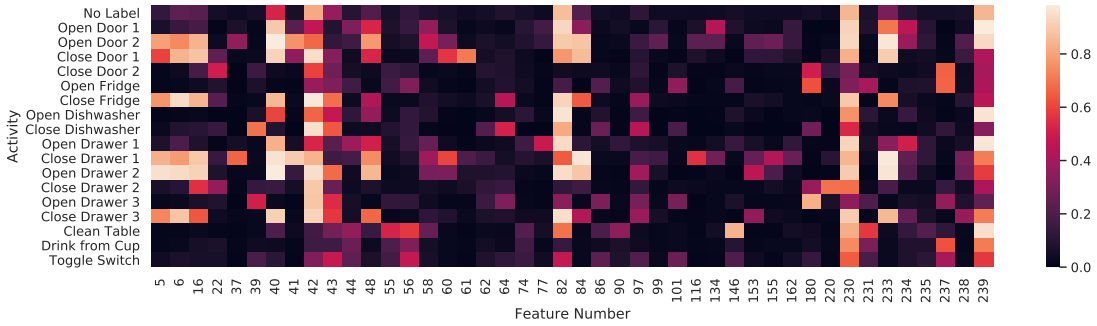


Figure 4: Heatmap of sensor feature activations under each activity of the OPPORTUNITY dataset. *Only active features are shown out of the 242 features in total.

(Roggen et al., 2010). This dataset consists of multiple different label types for human activity, ranging from locomotion, hand gestures, to object interactions. The dataset consists of 242 measurements from accelerometers and Inertial Measurement Units (IMUs) attached to the user, as well as accelerometers attached to different objects with which the user can interact. Additional experiment details can be found in the supplementary materials.

We use the mid-level gesture activities as the target for our models to predict, which contain gestures related to specific objects, such as opening a door and drinking from a cup. A comparison of the accuracy and the percentage of selected features by different models is given in Table 2, while example predictions and a trade-off curve are constructed and shown in Figures 3(a), 3(b), and 3(c), with a similar trend as the results on the UCI HAR dataset.

A heatmap for the selected features under each activity is shown in Figure 4. Here, the active sensor features across all activities are features 40 and 42, readings of the IMU attached to the subject's back, feature 82, readings from the IMU attached to the Left Upper Arm (LUA), and features 230 and 239, location tags that estimate the subject's position. We posit that these general sensor features are selected to track the subject's overall position and movements, as they are also predominantly selected in cases with no labels. Meanwhile, sensors 5, 6, and 16, readings from the accelerometer attached to the hip, LUA, and back, are specific to activities involving opening/closing doors or drawers.

NTU-RGB-D Dataset: We further test our proposed method on the NTU-RGB-D dataset (Shahroudy et al., 2016). This dataset consists of 60 different activities performed by either a single individual or two individuals. The measurements of this dataset are in the form of skeleton data consisting of 25 different 3D coordinates of the corresponding joints of the participating individuals. Additional experiment details can be found in the supplementary materials.

Table 3: Comparison of various methods for activity recognition on the NTU-RGB-D dataset. *Accuracy and average number of features selected are in (%).

Method	Accuracy (%)	Features (%)
No Selection	83.02	100
Soft Attention	83.28	100
Thresh. Attention	40.07	52.31
VFDS	**83.31**	**49.65**

We compare our method with three different baselines shown in Table 3: the baseline RNN architecture, soft attention, and thresholded attention baseline. We see that our method maintains a competitive accuracy compared to the baseline using less than 50% of the features. On the other hand, because the thresholded attention formulation is not specifically optimized for feature sparsity, we see that it performs significantly worse compared to the other methods. Meanwhile, the soft-attention slightly improves upon the accuracy of the base architecture. However, as also indicated by our other experiments, soft-attention is not a dynamic feature selection method, and tends to select 100% of the features at all times.

The results on these four datasets indicate that our dynamic monitoring framework provides the best trade-off between feature efficiency and accuracy, while the features that it dynamically selects are also interpretable and associated with the actual activity types.

4 CONCLUSIONS

We have introduced a novel method, VFDS, for performing foresight dynamic feature selection through variational inference. We accomplish this by defining a variational objective with a prior that captures the performance-cost trade-off of observing a given feature at a given time-point. We then designed an implicit,

covariate-dependent variational distribution, and use a differentiable relaxation, making the optimization amenable to stochastic gradient-based optimization. As our method is easily applicable to existing neural network architectures, we are able to apply our method on Recurrent Neural Networks for human activity recognition. We benchmark our model on four different activity recognition datasets and have compared it with various dynamic and static feature selection benchmarks. Our results show that our model maintains a desirable prediction performance using a significantly small fraction of the sensors or features. The features that our model selected were shown to be interpretable and associated with the activity types.

Acknowledgments

This project is supported in part by the Defense Advanced Research Projects Agency (DARPA) under grant FA8750-18-2-0027. X. Qian has been supported in part by the National Science Foundation (NSF) Awards 1553281, 1812641, 1835690, 1934904, and 2119103.

References

Anguita, D.; Ghio, A.; Oneto, L.; Parra, X.; and Reyes-Ortiz, J. L. 2013. A public domain dataset for human activity recognition using smartphones. In *21th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN)*, 437–442. CIACO.

Ardywibowo, R.; Huang, S.; Gui, S.; Xiao, C.; Cheng, Y.; Liu, J.; and Qian, X. 2018. Switching-state dynamical modeling of daily behavioral data. *Journal of Healthcare Informatics Research* 2(3):228–247.

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. In *The 22nd International Conference on Artificial Intelligence and Statistics (AISTATS 2019)*.

Ardywibowo, R.; Boluki, S.; Gong, X.; Wang, Z.; and Qian, X. 2020. NADS: Neural architecture distribution search for uncertainty awareness. In *International Conference on Machine Learning*, 356–366. PMLR.

Ardywibowo, R. 2017. *Analyzing Daily Behavioral Data for Personalized Health Management*. Ph.D. Dissertation.

Aziz, O.; Robinovitch, S. N.; and Park, E. J. 2016. Identifying the number and location of body worn sensors to accurately classify walking, transferring and sedentary activities. In *2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 5003–5006. IEEE.

Blei, D. M.; Kucukelbir, A.; and McAuliffe, J. D. 2017. Variational inference: A review for statisticians. *Journal of the American statistical Association* 112(518):859–877.

Bloom, V.; Argyriou, V.; and Makris, D. 2013. Dynamic feature selection for online action recognition. In *International Workshop on Human Behavior Understanding*.

Blundell, C.; Cornebise, J.; Kavukcuoglu, K.; and Wierstra, D. 2015. Weight uncertainty in neural network. In *International Conference on Machine Learning*, 1613–1622. PMLR.

Boluki, S.; Ardywibowo, R.; Dadaneh, S. Z.; Zhou, M.; and Qian, X. 2020. Learnable Bernoulli dropout for Bayesian deep learning. In Chiappa, S., and Calandra, R., eds., *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, 3905–3916. PMLR.

Chen, K.; Zhang, D.; Yao, L.; Guo, B.; Yu, Z.; and Liu, Y. 2020. Deep learning for sensor-based human activity recognition: Overview, challenges and opportunities. *arXiv preprint arXiv:2001.07416*.

Cheng, W.; Erfani, S.; Zhang, R.; and Kotagiri, R. 2018. Learning datum-wise sampling frequency for energy-efficient human activity recognition. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Cho, K.; Van Merriënboer, B.; Gulcehre, C.; Bahdanau, D.; Bougares, F.; Schwenk, H.; and Bengio, Y. 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*.

Ertuğrul, Ö. F., and Kaya, Y. 2017. Determining the optimal number of body-worn sensors for human activity recognition. *Soft Computing* 21(17):5053–5060.

Friedman, J.; Hastie, T.; and Tibshirani, R. 2008. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics* 9(3):432–441.

Friedman, J.; Hastie, T.; and Tibshirani, R. 2010. A note on the group lasso and a sparse group lasso. *arXiv preprint arXiv:1001.0736*.

Gal, Y., and Ghahramani, Z. 2016. Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In *International Conference on Machine Learning*, 1050–1059. PMLR.

Gal, Y.; Hron, J.; and Kendall, A. 2017. Concrete dropout. *Advances in Neural Information Processing Systems* 30.

Gordon, D.; Czerny, J.; Miyaki, T.; and Beigl, M. 2012. Energy-efficient activity recognition using prediction. In *International Symposium on Wearable Computers*. IEEE.

Graves, A.; Mohamed, A.-R.; and Hinton, G. 2013. Speech recognition with deep recurrent neural networks. In *2013 IEEE international conference on acoustics, speech and signal processing*, 6645–6649. IEEE.

Graves, A. 2011. Practical variational inference for neural networks. *Advances in neural information processing systems* 24.

Han, S.; Pool, J.; Tran, J.; and Dally, W. J. 2015. Learning both weights and connections for efficient neural networks. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'15, 1135–1143.

He, H., and Eisner, J. 2012. Cost-sensitive dynamic feature selection. In *ICML Inferning Workshop*.

Hinton, G.; Srivastava, N.; and Swersky, K. 2012. Neural networks for machine learning. *Coursera, video lectures* 264:1.

Hoffman, M. D.; Blei, D. M.; Wang, C.; and Paisley, J. 2013. Stochastic variational inference. *Journal of Machine Learning Research* 14(5).

Jang, E.; Gu, S.; and Poole, B. 2016. Categorical reparameterization with Gumbel-Softmax. *arXiv preprint arXiv:1611.01144*.

Jiang, Z.; Ardywibowo, R.; Samereh, A.; Evans, H. L.; Lober, W. B.; Chang, X.; Qian, X.; Wang, Z.; and Huang, S. 2019. A roadmap for automatic surgical site infection detection and evaluation using user-generated incision images. *Surgical infections* 20(7):555–565.

Karayev, S.; Fritz, M.; and Darrell, T. 2013. Dynamic feature selection for classification on a budget. In *International Conference on Machine Learning (ICML): Workshop on Prediction with Sequential Models*.

Kiefer, J. 1959. Optimum experimental designs. *Journal of the Royal Statistical Society: Series B (Methodological)* 21(2):272–304.

Kingma, D. P., and Welling, M. 2013. Auto-encoding variational Bayes. *arXiv preprint arXiv:1312.6114*.

Kingma, D. P.; Salimans, T.; and Welling, M. 2015. Variational dropout and the local reparameterization trick. *Advances in neural information processing systems* 28:2575–2583.

Kolamunna, H.; Hu, Y.; Perino, D.; Thilakarathna, K.; Makaroff, D.; Guan, X.; and Seneviratne, A. 2016. AFV: Enabling application function virtualization and scheduling in wearable networks. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, 981–991.

Koop, G., and Korobilis, D. 2018. Bayesian dynamic variable selection in high dimensions. *Econometrics: Econometric & Statistical Methods - Special Topics eJournal*.

Lei, T.; Barzilay, R.; and Jaakkola, T. 2016. Rationalizing neural predictions. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 107–117. Austin, Texas: Association for Computational Linguistics.

Li, C.; Chen, C.; Carlson, D.; and Carin, L. 2016a. Preconditioned stochastic gradient Langevin dynamics for deep neural networks. In *Thirtieth AAAI Conference on Artificial Intelligence*.

Li, J.; Chen, X.; Hovy, E.; and Jurafsky, D. 2016b. Visualizing and understanding neural models in NLP. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 681–691. San Diego, California: Association for Computational Linguistics.

Li, S.; Li, W.; Cook, C.; Zhu, C.; and Gao, Y. 2018. Independently recurrent neural network (IndRNN): Building a longer and deeper RNN. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 5457–5466.

Louizos, C., and Welling, M. 2017. Multiplicative normalizing flows for variational Bayesian neural networks. In *International Conference on Machine Learning*, 2218–2227. PMLR.

Louizos, C.; Welling, M.; and Kingma, D. P. 2017. Learning sparse neural networks through l_0 regularization. *arXiv preprint arXiv:1712.01312*.

MacKay, D. J. 1992. A practical Bayesian framework for backpropagation networks. *Neural computation* 4(3):448–472.

Maddison, C. J.; Mnih, A.; and Teh, Y. W. 2016. The concrete distribution: A continuous relaxation of discrete random variables. *arXiv preprint arXiv:1611.00712*.

Mahendran, A., and Vedaldi, A. 2015. Understanding deep image representations by inverting them. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Molchanov, D.; Kharitonov, V.; Sobolev, A.; and Vetrov, D. 2019. Doubly semi-implicit variational inference. In *The 22nd International Conference*

on Artificial Intelligence and Statistics, 2593–2602. PMLR.

Neal, R. M. 2012. *Bayesian learning for neural networks*, volume 118. Springer Science & Business Media.

Pierskalla, W. P., and Brailer, D. J. 1994. Applications of operations research in health care delivery. *Handbooks in operations research and management science* 6:469–505.

Ribeiro, M. T.; Singh, S.; and Guestrin, C. 2016. "Why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, 1135–1144. New York, NY, USA: Association for Computing Machinery.

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 *2010 Seventh international conference on networked sensing systems (INSS)*, 233–240. IEEE.

Satsangi, Y.; Whiteson, S.; and Oliehoek, F. A. 2015. Exploiting submodular value functions for faster dynamic sensor selection. In *AAAI Conference on Artificial Intelligence*.

Selvaraju, R. R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; and Batra, D. 2017. Grad-Cam: Visual explanations from deep networks via gradient-based localization. In *2017 IEEE International Conference on Computer Vision (ICCV)*, 618–626.

Shahroudy, A.; Liu, J.; Ng, T.-T.; and Wang, G. 2016. NTU RGB+D: A large scale dataset for 3d human activity analysis. In *IEEE Conference on Computer Vision and Pattern Recognition*.

Shen, X., and Varshney, P. K. 2013. Sensor selection based on generalized information gain for target tracking in large sensor networks. *IEEE Transactions on Signal Processing* 62(2):363–375.

Shi, J.; Sun, S.; and Zhu, J. 2018. Kernel implicit variational inference. In *International Conference on Learning Representations*.

Spaan, M. T., and Lima, P. U. 2009. A decision-theoretic approach to dynamic sensor selection in camera networks. In *Nineteenth International Conference on Automated Planning and Scheduling*.

Tartaglione, E.; Lepsøy, S.; Fiandrotti, A.; and Francini, G. 2018. Learning sparse neural networks via sensitivity-driven regularization. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, 3882–3892.

Tibshirani, R. 1996. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)* 58(1):267–288.

Titsias, M. K., and Ruiz, F. 2019. Unbiased implicit variational inference. In *The 22nd International Conference on Artificial Intelligence and Statistics*, 167–176. PMLR.

Vaizman, Y.; Ellis, K.; and Lanckriet, G. 2017. Recognizing detailed human context in the wild from smartphones and smartwatches. *IEEE Pervasive Computing* 16(4):62–74.

Welling, M., and Teh, Y. W. 2011. Bayesian learning via stochastic gradient Langevin dynamics. In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, 681–688. Citeseer.

Yang, X.; Chen, Y.; Yu, H.; Zhang, Y.; Lu, W.; and Sun, R. 2020. Instance-wise dynamic sensor selection for human activity recognition. *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI 2020)* 34(01):1104–1111.

Yin, M., and Zhou, M. 2018a. ARM: Augment-Reinforce-Merge gradient for discrete latent variable models. *arXiv preprint arXiv:1807.11143*.

Yin, M., and Zhou, M. 2018b. Semi-implicit variational inference. In *International Conference on Machine Learning*, 5660–5669. PMLR.

Zappi, P.; Lombriser, C.; Stiefmeier, T.; Farella, E.; Roggen, D.; Benini, L.; and Tröster, G. 2008. Activity recognition from on-body sensors: accuracy-power trade-off by dynamic sensor selection. In *European Conference on Wireless Sensor Networks*, 17–33. Springer.

Zou, H., and Hastie, T. 2005. Regularization and variable selection via the elastic net. *Journal of the royal statistical society: series B (statistical methodology)* 67(2):301–320.