Feature Learning for Enhanced Security in the Internet of Things

Enrico Mattei*, Cass Dalton*, Andrew Draganov*, Brent Marin
Michael Tinston, Greg Harrison, Bob Smarrelli, Marc Harlacher
Expedition Technology, Inc.

***Abstract*—Identifying Internet of Things (IoT) devices by their Radio Frequency (RF) fingerprint has important security implications. As the number of connected devices grows, current authentication mechanisms are becoming more susceptible to device spoofing attacks. To combat this, we exploit hardware imperfections in the RF transmit chain to extract device-specific features that uniquely identify an emitter, providing an additional layer of security. This is accomplished with a complex-valued Variational Autoencoder that has a Gaussian Mixture (GMVAE) prior on the latent variables' marginal distribution. By exploiting sequential information in the RF time-series data, we achieve processing gain by integrating multiple latent-space representations from a single device. We test and analyze the proposed approach on real WiFi data and obtain excellent classification results. We also test the proposed model on an Out-of-Distribution (OOD) detection task.**

***Index Terms*—Internet of Things, RF fingerprinting, Variational Inference, Deep Learning**

I. Introduction

The Internet of Things (IoT) is increasingly changing our lives. Connected devices are being incorporated into facets of both our work and home lives. However, millions of in-use IoT devices lack fundamental security measures and are open to attack. This leaves networks interacting with IoT devices vulnerable as well.

The ability to identify IoT devices based on their unique hardware-specific artifacts can mitigate many attacks that are commonly applied to insecure devices. For instance, hardware-specific authentication would prevent an identity spoofing man-in-the-middle attack. It can also be used to filter traffic from devices perpetrating a distributed denial of service (DDos) attack with compromised IoT devices. An attack of this sort successfully left much of the U.S east coast without internet access in 2016. The strength of such an authentication mechanism is that it assumes no software or hardware dependencies and hence can improve network security without any modifications to existing devices.

In an RF transmit chain, the digital in-phase (I) and quadrature (Q) signal components go through independent digital-to-analog converters (DACs), quadrature modulated, and amplified before being sent through the antenna. Each of these blocks imparts a signature on the transmitted signal that is specific to the device. For example, the DAC's input-output characteristics impose a nonlinear relationship known as the Integral Nonlinearity (INL), which measures the deviation between the ideal output value and the measured output value for a given input. Furthermore, the poles of the DAC reconstruction filters can deviate slightly from their nominal location due to component manufacturing tolerances. This deviation also introduces a device specific imperfection. Additionally, power amplifiers exhibit nonlinear behavior as input levels increase and may introduce in-band noise or interference in adjacent frequency channels. Lastly, the oscillators can introduce a phase imbalance in the data since there will always be a small phase offset between them causing the resulting I and Q channels to deviate slightly from perfect quadrature. All of these could be exploited for the purposes of device identification.

In this work, we model the device specific features as hidden random variables and propose an approach which exploits the expressive power of structured latent variable models in order to enforce a semantically meaningful latent space. The model is trained to learn location independent features in a supervised fashion from raw RF complex-valued data. In contrast to existing approaches which extract hand-crafted features, our approach is completely learned from data and therefore is agnostic to waveform and protocol specifics. We present results on datasets of 100 and 500 WiFi devices collected under normal use conditions. We show additional results on 19 WiFi devices that communicated bitwise identical data packets.

The next sections are organized as follows. In Section II we briefly review relevant work in RF fingerprinting. In Section III we describe the proposed model and justify its use based on knowledge of a traditional RF transmit chain. Experimental results are presented in Section IV, and we conclude in Section V along with providing future research directions.

II. Related work

In [1], it was proposed to model hardware imperfections with either a continuous Brownian Bridge process for the DAC or a Volterra series for the power amplifier. Device identification was then performed via standard hypothesis testing techniques. One drawback of this approach is that it has access to the decoded data and the inputs and outputs of each element of the transmit chain. This renders this approach impractical for most applications since it requires input-output measurements of individual components of each device that we wish to identify. Other works such as [2], [3], and [4] first

*Equal contribution

	Cplx Conv 1	Cplx Conv 2	Cplx Conv 3	Cplx Conv 4	Cplx Conv 5	Cplx FC 1	Cplx Conv FC 2	$q_{\phi(\boldsymbol{z}\|\boldsymbol{x})}$
#units	32	64	64	128	128	512	256	128
filter width	1	3	8	16	16	–	–	–
stride	1	1	3	8	8	–	–	–
activation	Cplx Cardioid	Cplx Cardioid	Cplx Cardioid	Cplx Cardioid	Cplx Cardioid	Cplx Cardioid	Cplx Cardioid	–

TABLE I: GMVAE encoder architecture

extract traditional RF features and a classifier or clustering algorithm is trained on those features. These approaches are usually designed for a specific communications protocol and signal modulation scheme. Deep learning based approaches are just beginning to be explored for the RF device identification problem, see [5] and [6], where complex-valued digitized RF signals are provided to a convolutional network architecture for classification. In [5], standard deep architectures are employed for device identification, but they require accurate carrier synchronization. The authors show results only for 7 ZigBee Pro devices. Learned deep probabilistic modeling has received little to no attention in the RF domain.

III. Generative RF Signal Modeling

A. The Canonical Model

The ideal signal transmitted by a wireless device can be expressed mathematically as

$$x(t) = Re\left[A\Big(h_{re} * a_{re}(t)\, cos\left(\omega(t)\right) + j h_{im} * a_{im}(t)\, sin\left(\omega(t)\right)\Big) e^{j2\pi f_c t}\right], \tag{1}$$

where h_{re} and h_{im} are the impulse responses of the in-phase and quadrature reconstruction filters that make up the DAC, $a(t)$ is the modulated digital signal amplitude, f_c is the carrier frequency, and A is the gain of the power amplifier. The complex IQ signal at the receiver side can be modeled as $r(t) = x(t) * h_c + \eta(t)$, where h_c is the complex-valued impulse response of the propagation channel, $*$ denotes the convolution operator, and $\eta(t)$ is a Gaussian noise component $\sim \mathcal{CN}(0, \sigma_n^2)$. As discussed in Section I, hardware imperfections cause the transmitted signal to deviate from its ideal representation. Since the DAC reconstruction filters are usually realized as relatively low-order analog filters, we can assume that the device specific features we are extracting occupy a low-dimensional subspace, according to the findings in [1]. In the sequel, we focus on learning a probabilistic model of these low-dimensional features from minimally pre-processed RF data.

B. Variational Autoencoder with Gaussian Mixture Prior

In general, the transmitter components that make devices (even from the same manufacturing lot) measurably different can be modeled as linear or nonlinear transformations of the inputs plus an additive noise component, $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{h} + \boldsymbol{\eta}$, where the component model $\boldsymbol{A} \in \mathcal{C}^{m \times d}$, is a matrix whose elements are functions of the input vector, and $\boldsymbol{h} \in \mathcal{C}^{d \times 1}$ is a vector which represents the component parameters of the device. In most applications we do not have access to the component's input and output signals to solve for the device parameters. To account for this, we leverage the knowledge that the transmitter components can be modeled by a small number of parameters and propose the following generative signal model to learn low dimensional features of RF devices,

$$y \sim cat\,(1/K) \tag{2a}$$
$$\boldsymbol{z} \sim \mathcal{CN}\left(\boldsymbol{\mu}_z(y), diag\left(\boldsymbol{\sigma}_z^2(y)\right)\right) \tag{2b}$$
$$\boldsymbol{x} \sim \mathcal{CN}\left(\boldsymbol{\mu}_x(\boldsymbol{z}), diag\left(\boldsymbol{\sigma}_x^2(\boldsymbol{z})\right)\right), \tag{2c}$$

where K is the number of devices, y the device id, $\boldsymbol{z}$ is the low-dimensional device representation, i.e., the latent variables, and $\boldsymbol{x}$ is the observed data. As stated previously, the generative process above implies that the marginal probability distribution of the latent variables is a Gaussian Mixture [7]. The posterior distribution, $p(\boldsymbol{z}, y|\boldsymbol{x})$, is usually intractable under a non-linear inference model such as a neural network. However, by using a variational approximation $q(\boldsymbol{z}, y|\boldsymbol{x})$ to the true posterior, the log marginal likelihood of $\boldsymbol{x}$ can be lower-bounded and subsequently optimized by the Evidence Lower Bound (ELBO),

$$\mathcal{L}(\theta, \phi; \boldsymbol{x}) = \mathbb{E}_{q_\phi(\boldsymbol{z}, y|\boldsymbol{x})}\left[\log p_\theta(\boldsymbol{x}|\boldsymbol{z})\right] - D_{KL}\left(q_\phi(\boldsymbol{z}, y|\boldsymbol{x}) \,||\, p(\boldsymbol{z}, y)\right), \tag{3}$$

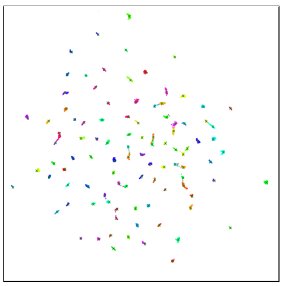
where $D_{KL}(q||p)$ is the Kullback-Liebler divergence between probability distributions q and p. We parameterize the inference network, $q_\phi(\boldsymbol{z}, y|\boldsymbol{x})$, and the generative network, $p_\theta(\boldsymbol{x}|\boldsymbol{z}, y)$, using neural networks, and learn their parameters by maximizing the ELBO and using the reparameterization trick [8]. In this work, we assume that device labels are available for training. The rationale for this assumption is due to the fact that the devices are nominally identical. Without any initial feature extraction, the model will fail to converge to a meaningful solution. Thus, once the device label has been observed, we optimize,

$$\mathcal{L}(\theta, \phi; \boldsymbol{x}) = \mathbb{E}_{q_\phi(\boldsymbol{z}|\boldsymbol{x})}\left[\log p_\theta(\boldsymbol{x}|\boldsymbol{z})\right] - D_{KL}\left(q_\phi(\boldsymbol{z}|\boldsymbol{x}) \,||\, p(\boldsymbol{z}|y)\right). \tag{4}$$

Note that $\boldsymbol{\mu}_{\boldsymbol{z}}(y)$ and $\boldsymbol{\sigma}_{\boldsymbol{z}}^2(y)$ are learned functions of the device label.

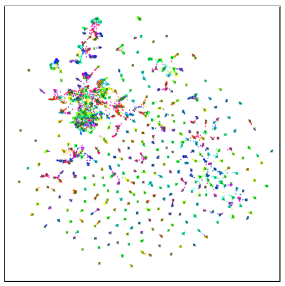
C. RF sequence processing

Since digitized RF signals can be viewed as a time series, we employ sequence processing techniques to obtain a significant processing gain. We achieve this by exploiting the fact that a single device will transmit several bursts of data while attempting to gain access to a network. Here, a burst refers



(a)

Accuracy: 96.4%
Std. Dev: 0.80%
5 cross-validation splits
Colors correspond to device label



(b)

Accuracy: 84.7%
Std. Dev: 0.473%
3 cross-validation splits
Colors correspond to device label

(c)

Out-Of-Library Detection Accuracy: 92.9%
False Positive Rate: 2.90%
cyan: detected out-of-library devices
yellow: missed out-of-library devices

Fig. 1: Umap plots on 100 device classification, 500 device classification and 90/10 out-of-library detection tasks.

to the transmission of a data message. Let $\boldsymbol{X} = \{\boldsymbol{x}_i\}_{i=1}^{M}$ be a sequence of M non-overlapping time windows of IQ data from the same device, where $\boldsymbol{x}_i \in \mathcal{C}^n$. Note that $\boldsymbol{X}$ is allowed to contain data from multiple bursts. The M latent vectors $\boldsymbol{z}_i$ should therefore be identical, and we denote this common latent vector as $\overline{\boldsymbol{z}}$. We can then model $\boldsymbol{z}_i|\boldsymbol{x}_i = \overline{\boldsymbol{z}} + \boldsymbol{\eta}_i$, where $\boldsymbol{\eta}_i$ is a zero-mean Gaussian noise component. That is, the latent variables of each sequence element are noisy observations of the sequence-level latents, $\overline{\boldsymbol{z}}$, representing the device. Assuming that the elements of the sequence are iid, we compute the sequence-level latent representation, as the maximum likelihood estimate of $\overline{\boldsymbol{z}}$, $\frac{1}{M}\sum_{i=1}^{M}\boldsymbol{z}_i$. The lower bound for the sequence becomes,

$$\mathcal{L}(\theta,\phi;\boldsymbol{X}) = \sum_{i=1}^{M}\mathbb{E}_{q_\phi(\overline{\boldsymbol{z}}|\boldsymbol{X})}\left[\log p_\theta(\boldsymbol{x_i}|\overline{\boldsymbol{z}})\right] - D_{KL}\left(q_\phi(\overline{\boldsymbol{z}}|\boldsymbol{X})\,||\,p_\theta(\boldsymbol{z}|y)\right). \quad (5)$$

In practice, this means that we process each burst in the sequence independently, then integrate the learned latent samples together by taking their mean.

Despite the truth labels, it is easy for the optimizer to get stuck in bad local minima. This is because we are trying to optimize $p_\theta(\boldsymbol{z}|y)$ jointly with $q_\phi(\overline{\boldsymbol{z}}|\boldsymbol{X})$. Therefore, we need to enforce samples drawn from the prior to be discriminative, while at the same time minimizing $D_{KL}\left(q_\phi(\overline{\boldsymbol{z}}|\boldsymbol{X})\,||\,p_\theta(\boldsymbol{z}|y)\right)$. To achieve this we add a small fully-connected classification head $p(y|\boldsymbol{z})$ with cross entropy loss that classifies inputs $\boldsymbol{z}$ that are sampled from $p(\boldsymbol{z}|y)$. This means that we are optimizing

$$\mathcal{L}_{total} = \mathcal{L}(\theta,\phi;\boldsymbol{X}) + \log p_\theta(y|\boldsymbol{z}). \quad (6)$$

IV. Experiments

We now present results showing the performance of our method using a dataset provided by the DARPA Radio Frequency Machine Learning Systems (RFMLS) program. Previous RF deep learning applications were often limited in scope due to the small datasets that they operated on. The RFMLS dataset, however, contains terabytes of labeled raw RF complex-valued data from thousands of IoT devices conforming to the IEEE 802.11a and 802.11g specifications, which allowed us to appropriately analyze our model and the results. Out of privacy concerns, the signal data is limited to contain only administrative packets.

Here, we present results on the task of identifying 100, and 500 devices collected "in the wild", with all the impairments and artifacts that are present in real world data. We also show the performance of our model against a high Signal-to-Noise Ratio (SNR) dataset of 19 devices, collected in a laboratory environment, for which the signals are identical bit for bit. No attempt was made to control the channel through which each device transmitted these bit-for-bit identical sequences. Additionally, we present results on an OOD detection task using a dataset of 100 devices where 90 devices were used to train the model and the remaining 10 devices were considered as OOD.

For each of these experiments, we employed an $80\%/20\%$ train/test split. The 100-item, 500-item and bitwise identical datasets respectively contain $109K$, $28K$, and $11K$ signals per device. The network architecture for the encoder is shown in Table I. In the encoder network all layers are complex-valued except the last layer which outputs the parameters of the approximate posterior. All complex-valued layers were implemented as described in [9]. The decoder network consists simply of three complex-valued fully-connected layers. In all experiments, the model was trained using the RMSProp optimizer with a mini-batch size of 64, a 0.0005 learning rate and no gradient clipping. We used the complex cardioid activation function [10] in all complex-valued layers. Our models were trained for $100,000$ steps. To obtain the classification results we use $p_\theta(y|\boldsymbol{z})$ defined in Equation (6) as our classifier. Some hardware components can provide a signature that could be used for classification, such as the carrier frequency offset. However, these signatures can vary slowly over time as a function of the physical environment. Such time-varying signatures would prevent us from correctly identifying a particular device if it has drifted at inference time. To mitigate this we perform

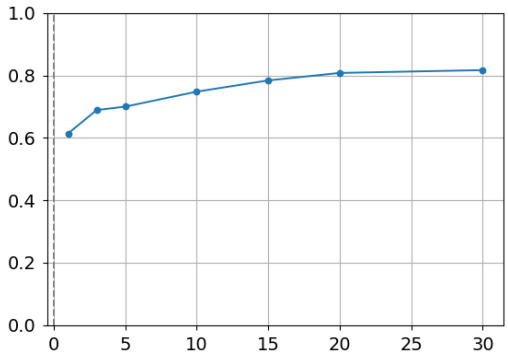


Fig. 2: 100 in-the-wild devices. Evaluation set accuracy as a function of the sequence length M.

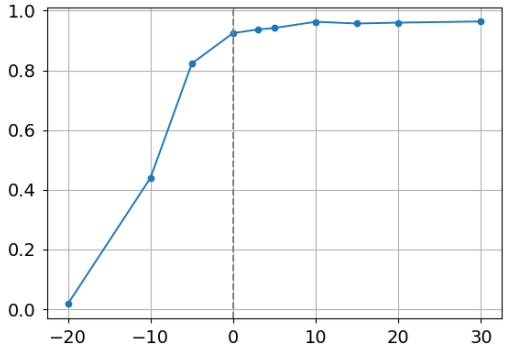


Fig. 3: 100 in-the-wild devices. Evaluation set accuracy as a function of SNR.

data augmentation during training, where random frequency offsets are applied to the signals prior to being consumed by the network. This augmentation prevents the model from learning these signatures as discriminative features.

A. Classification Results

The proposed model achieves excellent results in the traditional classification setting on both the 100 and 500 device datasets. We cross-validated the results on multiple random train/test splits and list the results in Figures 1a and 1b. Furthermore, we conducted an additional test to assert that our model is learning features that are not correlated with the device's MAC address. Specifically, we trained the model on a dataset of 19 devices with the same MAC address and bitwise-identical signals and obtained 99.8% classification accuracy on the test set. This verifies the assumption that the model can learn discriminative features that are intrinsic to the physical device.

We demonstrate the advantage of our sequence-processing approach in Figure 2. To perform this test, we generated seven different instances of the 100 device dataset with varying sequence of lengths. Each dataset was made with $1K$ signals per device. This means that there are thirty times fewer sequences per device in the $M = 30$ dataset than there are in the $M = 1$ dataset. The results confirm that operating on multiple signals simultaneously helps to consolidate the hardware's unique fingerprint. Note that the accuracy is lower for this experiment due to the smaller number of training samples as the sequence length grows.

Results from training the model on signals with random SNR values and evaluating on signals with set SNR in are shown in Figure 3. We artificially imposed the SNR by assuming that the burst is entirely comprised of the relevant signal and then elevating the noise floor to reach a desired signal-to-noise ratio. Due to assuming that the signals are noiseless, the reported SNR is an upper bound on the true signal-to-noise ratio of the data.

B. Out of Distribution Detection

We now test our ability to detect OOD devices using the proposed model. For this, we created a dataset by removing 10 of the devices from both the train and test splits of the 100-device classification dataset from Figure 1a to test if our model could identify the signals from these 10 devices as being OOD. We then trained a model on the remaining 90 devices and tested for in-distribution and out-of-distribution detection by operating on the union of the test-set devices and out-of-library devices, we refer to this union as the "inference" device set.

In order to detect whether a signal in the inference set is not from the training distribution, we exploit the probabilistic nature of our model as follows: we classify the inference signals using the trained model, then we perform dimensionality reduction, using UMAP [11], on the latent representations of the inference signals as well as the training set signals and the mean vectors of the model's GMM. Once we have obtained the low-dimensional projections, we perform a simple Z-Test on the UMAP-space distances from each of the latent representations in the inference set to its assigned Gaussian cluster. Using this procedure, our model correctly detected 92.9% of the out-of-library signals in the $90/10$ dataset. The false positive rate of test-set devices that got detected as out-of-library was 2.9%. The latent representation can be seen in Figure 1c. Note that the new devices land in distinguishable clusters.

V. Conclusions and Future Work

We have presented a sequence based probabilistic model for classification of IoT devices. It imposes a Gaussian Mixture prior on the sequence-level latent variables. We have shown excellent classification accuracy on real datasets of 100 and 500 devices. We analyzed these results with respect to relevant data variations and asserted that the learned features are independent of the signal payloads. Furthermore, we have shown that the model can recognize OOD devices with high accuracy and a low false positive rate. Future work includes leveraging the probabilistic foundations of the model to investigate larger population sizes (over 1000 devices), and performing low-shot learning on detected OOD devices.

ACKNOWLEDGEMENT

This research was developed with funding from the Defense Advanced Research Projects Agency (DARPA). The views, opinions and/or findings expressed are those of the authors and should not be interpreted as representing the official views or policies of the Department of Defense or the U.S. Government.

REFERENCES

[1] Adam C Polak, Sepideh Dolatshahi, and Dennis L Goeckel, "Identifying wireless users via transmitter imperfections," *IEEE Journal on selected areas in communications*, vol. 29, no. 7, pp. 1469–1479, 2011.

[2] S. V. Radhakrishnan, A. S. Uluagac, and R. Beyah, "Gtid: A technique for physical deviceanddevice type fingerprinting," *IEEE Transactions on Dependable and Secure Computing*, vol. 12, no. 5, pp. 519–532, Sept. 2015.

[3] Ke Gao, C. Corbett, and R. Beyah, "A passive approach to wireless device fingerprinting," in *Proc. IEEE/IFIP Int. Conf. Dependable Systems Networks (DSN)*, June 2010, pp. 383–392.

[4] N. T. Nguyen, G. Zheng, Z. Han, and R. Zheng, "Device fingerprinting to enhance wireless security using nonparametric bayesian method," in *Proc. IEEE INFOCOM 2011*, Apr. 2011, pp. 1404–1412.

[5] K. Merchant, S. Revay, G. Stantchev, and B. Nousain, "Deep learning for RF device fingerprinting in cognitive communication networks," *IEEE Journal of Selected Topics in Signal Processing*, vol. 12, no. 1, pp. 160–167, Feb. 2018.

[6] S. Riyaz, K. Sankhe, S. Ioannidis, and K. Chowdhury, "Deep learning convolutional neural networks for radio identification," *IEEE Communications Magazine*, vol. 56, no. 9, pp. 146–152, Sept. 2018.

[7] Zhuxi Jiang, Yin Zheng, Huachun Tan, Bangsheng Tang, and Hanning Zhou, "Variational deep embedding: An unsupervised and generative approach to clustering," *arXiv preprint arXiv:1611.05148*, 2016.

[8] *Auto-encoding variational bayes*, 2013.

[9] Chiheb Trabelsi, Olexa Bilaniuk, Ying Zhang, Dmitriy Serdyuk, Sandeep Subramanian, João Felipe Santos, Soroush Mehri, Negar Rostamzadeh, Yoshua Bengio, and Christopher J. Pal, "Deep Complex Networks," *arXiv:1705.09792 [cs]*, May 2017, arXiv: 1705.09792.

[10] P. Virtue, S. X. Yu, and M. Lustig, "Better than real: Complex-valued neural nets for mri fingerprinting," in *Proc. IEEE Int. Conf. Image Processing (ICIP)*, Sept. 2017, pp. 3953–3957.

[11] L. McInnes, J. Healy, N. Saul, and L. Grossberger, "Umap: Uniform manifold approximation and projection," *arXiv preprint arXiv:1802.03426*, 2018.