Wideband Spectral Monitoring Using Deep Learning

Horacio Franco
SRI International
Menlo Park, CA, USA
horacio.franco@sri.com

Chris Cobo-Kroenke
SRI International
Menlo Park, CA, USA
chris.cobo-kroenke@sri.com

Stephanie Welch
SRI International
Menlo Park, CA, USA
steph.welch@sri.com

Martin Graciarena
SRI International
Menlo Park, CA, USA
martin.graciarena@sri.com

ABSTRACT

We present a system to perform spectral monitoring of a wide band of 666.5 MHz, located within a range of 6 GHz of Radio Frequency (RF) bandwidth, using state-of-the-art deep learning approaches. The system detects, labels, and localizes in time and frequency signals of interest (SOIs) against a background of wideband RF activity. We apply a hierarchical approach. At the lower level we use a sweeping window to analyze a wideband spectrogram, which is input to a deep convolutional network that estimates local probabilities for the presence of SOIs for each position of the window. In a subsequent, higher-level processing step, these local frame probability estimates are integrated over larger two-dimensional regions that are hypothesized by a second neural network, a region proposal network, adapted from object localization in image processing. The integrated segmental probability scores are used to detect SOIs in the hypothesized spectro-temporal regions.

CCS CONCEPTS

• Networks → Network components → Wireless access points, base stations and infrastructure → Cognitive radios

KEYWORDS

Deep learning, wideband spectral monitoring, cognitive radio

ACM Reference format:

Horacio Franco, Chris Cobo-Kroenke, Stephanie Welch and Martin Graciarena. 2020. Wideband Spectral Monitoring Using Deep Learning. In *ACM Workshop in Wireless Security and Machine Learning (WiseML2020). July 13th, 2020, Linz, Austria,* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3395352.3402620

1 INTRODUCTION

Deep neural network architectures [1], have in recent years produced large performance improvements in image and speech recognition compared with more traditional techniques [2, 3]. Instead of using carefully handcrafted features with elaborate statistical modeling, deep learning approaches take as their input raw two-dimensional images, or simple spectro-temporal representations of audio signals and process them through multiple layers of neural computation, such as convolutional, fully connected, and recurrent layers. Their internal representations can be interpreted as powerful feature extractors. Multiple layers progressively combine features and derive abstractions, achieving high robustness to input variability when large amounts of representative training data are used.

Deep learning techniques have recently been applied in the RF domain [4, 5, 6]. Most of these systems deal with classifying the type of modulation of a single RF signal that has been located in frequency and shifted to the baseband.

In this paper we address the challenging problem of detection, identification, and localization in time and frequency of multiple SOIs in a background of RF activity that may include many other signals in a very wide band of 666.5 MHz, located anywhere between 0.5 and 6 GHz. This problem is part of the Spectrum Awareness task in the DARPA Radio Frequency Machine Learning Systems (RFMLS) program [7].

Current software-defined radios (SDRs) allow the digitization of over 600 MHz of RF bandwidth anywhere in the RF spectrum up to several GHz. One of the challenges in applying machine learning (ML) approaches to detection of a variety of RF signals is the huge range of bandwidths and durations across different signal types. This range exceeds four orders of magnitude, and no similar range across different signals or images is found in machine learning applications in speech or video. Other challenge is the large number of samples in digitized RF signals in a wide spectral band: with a sampling rate of 666.5 MHz, each second of RF IQ signals requires over 2.6 GB of storage. Thus, the processing pipeline must handle massive data quantities.

2 SYSTEM DESCRIPTION

Given the IQ output of the SDR the first step is to compute a high-resolution spectrogram of the 666.5 MHz bandwidth of interest. This spectrogram can be thought of as a two-dimensional "image" with frequency along one axis and time along the other. In this two-dimensional representation, signals can appear at any frequency location and bandwidth, and in any time interval. Without loss of generality, and for practical reasons, the

WiseML '20, July 13, 2020, Linz (Virtual Event), Austria

ACM ISBN 978-1-4503-8007-2/20/07…$15.00
https://doi.org/10.1145/3395352.3402620

continuous IQ output stream from the SDR is divided into segments of approximately 0.75 seconds duration. Each segment corresponds to ~500 million complex samples, or ~2 GB.

To generate the input signal spectrogram, we compute a sequence of non-overlapping 16K-point FFTs, apply a temporal Blackman window, obtain the power spectrum, and average N successive power spectra to reduce the feature dimensionality and reduce the variance of the power spectrum estimate. Finally, the log of the power spectrum is used for the spectrogram representation. With N = 2 the size of the spectrogram image is 16,384 bins in the frequency dimension and 15,360 log-power spectral frames in the time dimension. The resulting spectrogram "image" has over 250 million pixels.

The dimension of the spectrogram is too large to process with a single network of reasonable dimensions. We apply a hierarchical approach where at the lower level we analyze the spectrogram using a sweeping spectro-temporal window, or spectrogram frame. The spectrogram frame is input to a deep convolutional network that computes local probability estimates for the presence of SOIs within each frame. In a subsequent, higher-level stage, these local frame probability estimates are integrated over larger two-dimensional regions, or "bounding boxes", that are hypothesized by a second neural network: a "region proposal network" [8, 9]. Using the spectrogram frames within the bounding boxes we can obtain two-dimensional segmental probability scores for the presence of SOIs over a spectro-temporal range greater than a single frame. This approach allows the modeling of complex spectro-temporal behavior of signals that present wide frequency excursions and extended temporal activity patterns. The SOI segmental probability scores are used to detect the presence of SOIs. To estimate the frequency location for some very narrowband SOIs we introduced an optional processing step to refine the bounding boxes within the frames' bandwidth. A block diagram of the complete system is shown in Fig. 1.

2.1 SOI frame probability estimator

In this section we describe the feature extraction and the deep neural network (DNN) that computes frame-level SOI probability estimates. The input to this network is obtained by dividing the input spectrogram into frames of 256 by 256 bins in the frequency and time domains respectively; these frames overlap by 50% on the frequency dimension. The frames have a bandwidth of ~10.4 MHz and a duration of 12.6 msec. For each spectrogram frame, the DNN estimates a set of posterior probabilities for the presence of each SOI within that frame. We use a state-of-the-art deep convolutional neural network (DCNN) to estimate the SOI posterior probabilities, as well as a background signal class. The frame duration allows the network to capture temporal structure that a signal may show within the frame, while longer time structure is captured by the subsequent higher-level processing.

While the frequency bandwidth of ~10 MHz allows capture of a large number of signals, to better capture the structure of wider band signals we derive a multiscale representation from the same input spectrogram. By iteratively integrating across adjacent power spectral bins, we generate input frames with 2, 4, and 8 times the bandwidth of the basic frame while keeping the same 256 by 256 dimensionality for these wider-band frames. The bandwidths of the multiscale frames are 20.81, 41.62, and 83.25 MHz, respectively.

Frames corresponding to every bandwidth are input to different DCNNs, one for each bandwidth. The outputs of these nets are combined by averaging their values across all bandwidths, for every frame location, for each SOI. These combinations are done at the finest, 10.4 MHz, level of resolution. The aim of this output combination is to obtain a more robust estimate of the posterior probabilities for each SOI class across a range of bandwidths for every spectro-temporal frame.

We can consider the output of this first processing stage as converting the very large two-dimensional input spectrogram (of 250 million pixels) into a much smaller three-dimensional "posteriogram" (with ~84K elements at 50% frame overlap in frequency) that contains for every frame position, indexed by time and frequency location, a set of M smoothed posterior probabilities. The output of the "frame posterior estimator" is a tensor with 127 columns along the frequency dimension, 60 rows across the temporal dimension, and a third dimension with M = 11 values of posterior probabilities, corresponding to the set of 10 SOIs plus a background class, for each spectro-temporal cell position.

The architecture of the frame posterior estimation network, depicted in Fig. 2, is an extension of the 50-layer Residual Network (ResNet) proposed by He et al. [10, 11], including the architectural improvements reported in [12]. To reduce the input dimensionality, we added three initial convolutional layers, with 32, 32, and 64 convolution filters each, using 3x3 kernels, and 2x2 max pooling between layers. These are followed by four stages of residual blocks with 3, 4, 6 and 3 blocks each. At the end of every stage the feature maps' spatial dimension is halved and the number of features extractors is doubled, resulting in 64, 128, 256,

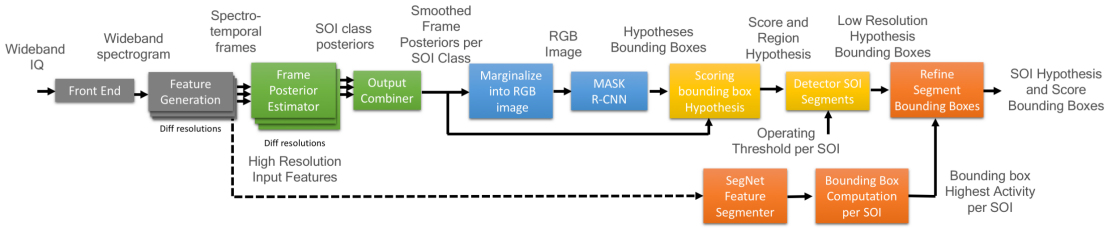

Figure 1: Block diagram of the proposed architecture for wideband spectral monitoring

and 512 feature maps respectively. That is followed by a Global Average Pooling (GAP) layer, pooling over spatial dimensions by averaging, and again by maxing. The GAP output of 512+512 units is followed by two fully connected layers of 512 units and a softmax output layer with 11 outputs.

This extended ResNet was trained from scratch on generated RF data. The training data size was ~1.27 million feature frames with SOIs randomly located within each frame. The training frames were extracted from wideband spectrograms obtained from generated wideband IQ data using real RF noisy backgrounds. The training data was selected to balance class priors before training. We doubled the training data size by on-the-fly data augmentation: we used random crops, frequency shifts, and added mix-up augmentation by weighted combinations (0.9, 0.1) of feature frames, which adds robustness to RF signal collisions. Four different networks were trained corresponding to each of the four frame bandwidths. Average classification accuracy for ten SOIs (from Table 1) plus background, on the noisy validation set, described in Section 4, was ~75.%.

2.2 Bounding box estimation and detection score computation

The following, higher-level, processing stage takes the posteriograms generated by the frame probability estimator ResNet, consisting of smoothed posterior probabilities for every SOI and every frame, and uses them to hypothesize a "bounding box", across a rectangular set of frames, for every potential SOI detection. This rectangular bounding box per SOI potential detection is aimed to capture a "human-level" interpretation of the spread of an RF signal along the spectral and temporal dimensions, and is part of the information provided by the spectral monitoring system after detection of an SOI.

Recent advances in image processing have successfully tackled the problem of bounding box estimation for objects present in a two-dimensional image. These systems, such as the Mask R-CNN [8], have been trained with millions of labeled and segmented images, and achieve very accurate object detection and bounding box estimation performance on standard image datasets. These systems, and their corresponding trained models, are publicly available [9], and are known to be adaptable to new tasks, by only re-training a small subset of the parameters of their deep neural networks. The re-training is done using a much smaller data set than was originally used to train the entirety of the deep neural network.

This limited re-training for a new task, referred to as transfer learning, is known to be effective in dealing with new types of objects in pictures. We apply this concept of transfer learning to bounding box estimation for SOIs in an RGB "image" derived from the posteriogram representation.

To allow transfer learning for the Mask R-CNN from image processing to posteriogram processing, we derive a three-channel image from the three-dimensional posteriogram by marginalizing the M posterior probabilities for each frame over three sets of SOI classes: wideband, narrowband, and background. In this way the marginalized posteriogram can be considered as a three-channel image with 127 pixels along the frequency axis (using 50% frame overlap) and 60 pixels along the time axis. This derived image is processed by the Mask R-CNN to estimate candidate bounding boxes for SOIs present in each 0.75-sec RF signal segment.

The resolution of the bounding box estimate is limited by the size of the spectro-temporal frames, 10.4 MHz in frequency and 12.6 msec. in time. This level of resolution is sufficient for detection and location of a large number of SOIs with reasonable approximation. For the few SOIs that require greater resolution to precisely locate them in the spectro-temporal space, like narrowband signals for voice communications, we have implemented a complementary segmentation processing that locates such signals within the spectrogram frames.

The Mask R-CNN also produces a score representing the goodness of match of the input data within the bounding box to the re-trained Mask R-CNN model. This score is combined with the SOI frame posterior scores within the bounding box to produce a segmental score for each SOI hypothesis within each bounding box hypothesis. To detect an SOI, its segmental score is compared with a calibrated threshold that is specific to each SOI, and a positive detection occurs when the score is above the threshold. We choose a threshold per SOI to obtain a desired tradeoff between false alarms and correct detections. To determine the thresholds, we plot ROC curves by sweeping the threshold over a range of values and pick the desired operating point in a development data set.

The Mask R-CNN architecture is summarized here following [8]: It has three main components: i) a convolutional backbone (a ResNet), ii) a Region Proposal Network, and iii) Bounding Box Regression and Classification heads. The convolutional backbone is a ResNet pretrained on the ImageNet database (1K classes, ~14 million frames) to do object classification. The system produces as output the feature maps from several layers with different

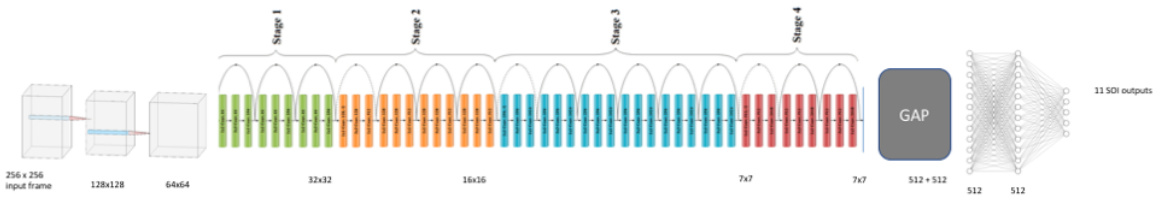

Figure 2: Frame Posterior Network architecture, based on ResNet 50 with added input convolutional layers, a GAP layer, and two fully connected layers to the softmax output layer

resolutions and information. The Region Proposal Network component also uses convolutional layers that find candidate SOI bounding boxes. Each candidate region selects feature subsets from the backbone feature maps and transforms them into a uniform shape (7x7) feature map. Finally, the classification and regressions heads process the potential regions to do final object classifications and produce bounding box coordinates, as well as a pixel-level mask for the detected object. Using as ground truth the known bounding boxes from our inserted SOIs we trained only the last layers: the region proposal network, the regression net, and the mask net. The pre-trained backbone convolutional net was fixed. About ~8000 RGB images were used for re-training, obtained from the posteriograms computed from all available wideband training data (with size 16 TB) by the frame posterior estimation networks.

The segmental score for an SOI hypothesis was computed as a weighted combination of two components: 1) the sum of the frame log-probabilities for the SOI class over all frames in the bounding box, and 2) the corresponding SOI detection score from the Mask R-CNN classification layer. The first component captures a robust segmental probability score for the SOI class, while the second component represents the level of match of the image to the patterns learned by the Mask R-CNN for that SOI.

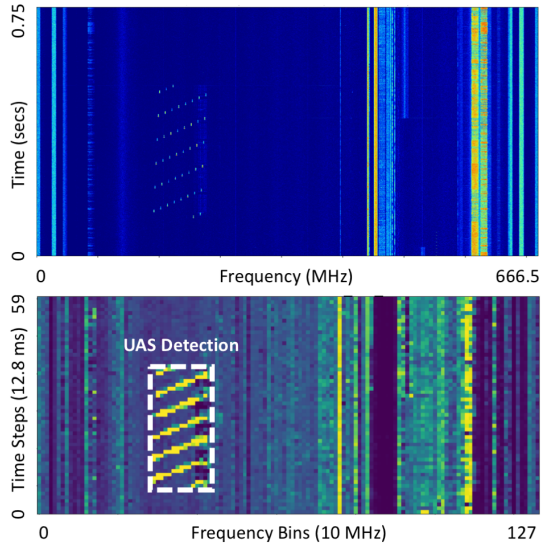

Figure 2: Top: Spectrogram showing drone RF activity, Bottom: Posteriogram slice for drone class, and estimated SOI detection bounding box

In Fig. 3 we show an example of segmental detection for a drone SOI. The top pane shows an input wideband spectrogram of 0.75 sec. duration and 666.5 MHz bandwidth. At left of center we can see a drone signal (UAS) with a periodic wideband pattern of RF activity, moving downward in frequency. In the second pane we show one "slice" of the posteriogram tensor for the class UAS (drone). We can clearly see the yellow downward stripes associated with the drone activity at the center left. Note that there are other regions that show similar probability values for signals that are not drones (vertical yellow lines in the right half of the middle pane). We also show the estimated bounding box for the detected drone SOI. Despite some high probability frames for the drone class on the right side, the time-frequency pattern structure of that RF activity does not match the "image" patterns for drones learned by the Mask R-CNN. Note that the background signals on the right side in the first pane are much stronger than the inserted drone signal. This is a typical case, as our insertions have positive signal to noise ratio (SNR) only with respect to the background noise floor.

To refine the estimate of bounding boxes for very narrowband signals, we implemented an optional processing stage that uses an image segmentation approach to locate SOI activity within the frames that intersect the low-resolution bounding box edges obtained from the posteriogram-derived image by the previous stage. Our image segmentation approach was based on the SegNet architecture [13] and consists of an encoder-decoder DNN with 500K trainable parameters. The output of the SegNet has the same dimensions as the input frame, with each pixel classified as the highest scoring class. We define the area of detection by selecting the largest contiguous area corresponding to the SOI detected in the analyzed frame. We trained the SegNet from scratch using internally generated data. We used 1.2 million training frames leveraging the existing training set for the frame posterior estimation net. The high-resolution annotations from an energy detector were used to generate target class labels for every pixel. Having obtained the regions of activity within each frame in the bounding box edges, we simply adjusted the overall SOI bounding box location according to the finer level of detail of SOI activity within the frames provided by the SegNet.

3 DATA GENERATION AND MODEL TRAINING

In the field of speech recognition there has been a long research effort to build models with robustness to background noise, channel conditions, and signal variability. The combination of multi-condition (MC) training and deep neural network models has been shown to be highly effective, and to have similar performance to other, more complicated, state-of-the-art approaches [14]. We apply and extend MC training to RF signals to train robust models in the RF domain by generating training data with considerable variability in several dimensions: i) multiple examples of each SOI, representing different signal behaviors, ii) examples of a range of signal-to-noise ratios, iii) examples with multiple SOI locations in time and frequency within the wideband spectrogram and within the spectrogram frames, iv) examples against different backgrounds and at different locations within the backgrounds.

The RFMLS program evaluation and development data was produced by a synthetic data generation process. The RFMLS program provided wideband recordings of multiple RF background environments: rural, urban, and operational. The length of each background environment recording was 120

seconds. To keep IQ files within manageable sizes, each background recording was segmented into 0.75-sec segments. The RFMLS program also provided baseband recordings of multiple signals (the potential SOIs) recorded over the air or over a wire. Development and evaluation RF signals were then created by inserting SOIs at specified times and frequencies over the given backgrounds, at specified SNRs with respect to the noise floor of the backgrounds. In Table 1 we show a list of the types of signals inserted, and the average bandwidth estimated by the energy detector as well as the typical bandwidth per channel.

The SOI insertions were implemented by first upsampling the baseband SOIs to the 666.5 MHz sampling rate of the wideband backgrounds, and then frequency-shifting the SOIs to the desired center frequencies. The specified SNRs were used to determine the amplitude of each inserted SOI.

We also generated a large amount of synthetic RF data to train the ML models. The training data was generated to produce a wide range of variability in terms of SOI samples, insertion locations, insertion SNRs, and background recordings. On average, 500 samples with SOIs were created and inserted in each 120 sec background. The SNRs of the SOI insertions ranged from 5 to 25 dB. When inserting SOI's we aimed to avoid overlap with preexisting signal activity in the recorded backgrounds.

We generated a total of 50 different training environments of 120 seconds each. The total generated IQ training data amounted to ~16 terabytes. From this training data set we extracted 1.27 million feature frames across all SOIs insertions to train the ResNet frame posterior probability estimators. To train the Mask R-CNN for bounding box estimation we generated ~8000 RGB images from the posteriograms derived from all the training data. Each training RGB image of 127 by 60 pixels comes from a 0.75-sec segment of IQ data, 2GB in size. There are relatively few training images, as they are very expensive to obtain, but as the Mask R-CNN is pre-trained, this data is only used to re-train a subset of its parameters.

In addition to SOI class labels, the evaluation and training data should include the locations, in terms of bounding boxes, of each SOI in time and frequency. This information was created, for each SOI sample, from the original baseband clean versions of the SOIs before insertion in the backgrounds. We used an energy detector software customized for each type of SOI that was provided to the RFMLS performers. This software produced two types of annotations per signal: 1) high-resolution annotations indicating where there is SOI energy, and 2) overall activity bounding box annotations surrounding the detailed activity for each SOI baseband recording. Both types of annotations were translated in frequency and time when an SOI was frequency shifted and inserted into a background. The high-resolution annotations were used to provide class labels for the SOIs within each training frame for the frame posterior estimation networks, and also used to train the bounding box refinement SegNet, while the overall activity bounding boxes were used to train the Mask R-CNN to predict the SOI bounding boxes.

Table 1: Signals inserted as potential SOIs

Class	Description	Average Bandwidth (Energy Detector)	Typical Protocol Bandwidth Per Channel
ADSB	Aircraft Positional beacon	53 MHz	50KHz or 1.3MHz
APCO-25	Emergency services digitial communication	12.5 KHz	12.5 KHz
ATSC	Digital Television Protocol	13 MHz	6MHz
Bluetooth	Short range data exchange protocol.	~1 MHz	1MHz
DECT	Cordless audio devices such as phones	5.6 MHz	1MHz
LTE	Wireless broadband communication protocol	23 MHz	1.4-20MHz
UAS	Collection of drone up/down link communications	80 MHz	Varies
VULOS	Collection of modulations for analog voice and data	50 KHz	Up to 25 KHz
WIFI	Common wireless network protocol	138 MHz	20-80 MHz
WCDMA	Mobile telecommunications protocol	5 MHz	5 MHz
WIMAX	Wireless broadband communication protocol	7.5 MHz	1.2-20MHz

4 EXPERIMENTAL RESULTS

In initial testing and optimization of the SOI detection approach, we defined the thresholds for detection of each class of SOI on a development set consisting of a 120-sec background segment where we inserted 423 SOIs using SNRs between 5 and 25 dB.

The evaluation sets were constructed by inserting specified signals at 10 dB SNR using a different background of 120 sec that was not used for generating training data. Tables 2 through 5 present the SOI detection performance for four RF evaluation environments with different signal insertion densities. Two full segments of 120 sec from each of the four RF environments were generated. The second and third RF environments are progressively more crowded with additional signals, and consequently more challenging. The fourth environment is closer to the first one in complexity, and shows results for other SOIs.

A detection, consisting of an SOI label and the associated bounding box, was considered to be correct if the area of the intersection between the ground truth bounding box and the detected bounding box, divided by the area of the detected bounding box, exceeded 50%.

We show performance results, as percent of correct detections and of false alarms, for a subset of the signals inserted into the backgrounds: DECT, Bluetooth (BT), ATSC, LTE, and drones (UAS) in the first and second environment; in the third test environment LTE is replaced by Vulos. In the fourth test environment we added ADSB and WiFi as new SOIs, keeping LTE, BT and UAS. The remaining inserted signals are considered distractors for these evaluations. We obtained very good detection performance for BT of 84% to 98% in all environments, and quite good (75% to 100%) for drones in all but one test set (51%). Performance for DECT was reasonably good, at above 70% detection, while the detection performance for ATSC was marginally acceptable at around 50%. For ATSC it seems the lower accuracy is due to poor performance of the energy detector in generating training bounding boxes. For LTE our development performance was much better than what we see in Environments 1 and 2. It seems there was a mismatch with those evaluation sets due to the use of some level of filtering that was not present in the training data. That was fixed in Env. 4, were we got 70% and 85% detection rates. Vulos is a very narrowband signal, and the only one in the evaluation set that required the additional SegNet processing to refine the bounding boxes. While the bounding box

refinement allowed a certain level of detection for this class of signals, the detection performance was poorer than for the other SOIs. WiFi showed very good detection performance at 92% and 89%. ADSB detection performance was also very good, but with a higher FA rate than on other signals.

Table 2: Detection performance in Environment 1

		DECT	BT	ATSC	LTE	UAS
Env. 1	Det %	76	88	50	50	100
Seg. 1	FA %	49	21	51	63	39
Env. 1	Det %	72	92	48	54	96
Seg. 2	FA %	38	19	59	60	27

Table 3: Detection performance in Environment 2

		DECT	BT	ATSC	LTE	UAS
Env. 2	Det %	71	87	44	57	82
Seg. 1	FA %	39	7	57	61	44
Env. 2	Det %	82	98	53	41	51
Seg. 2	FA %	32	14	46	69	41

Table 4: Detection performance in Environment 3

		DECT	BT	ATSC	Vulos	UAS
Env. 3	Det %	72	84	49	35	84
Seg. 1	FA %	27	3	50	50	47
Env. 3	Det %	70	89	51	28	80
Seg. 2	FA %	33	23	58	61	38

Table 5: Detection performance in Environment 4

		ADSB	BT	WIFI	LTE	UAS
Env. 4	Det %	97	89	92	70	83
Seg. 1	FA %	58	20	30	30	38
Env. 4	Det %	100	86	89	85	75
Seg. 2	FA %	42	20	35	6	28

5 CONCLUSION

We have presented a system for wideband spectral monitoring, i.e. detection of SOIs, in a 666.5 MHz-wide signal. We proposed an architecture to deal with the extremely large range of signal bandwidths and durations, as well as the data processing capabilities required to handle signals at the required sampling rate. We incorporated state-of-the-art deep learning techniques inspired by image processing, and segmental scoring techniques inspired by speech processing that were extended to the two-dimensional spectro-temporal space where RF signals are represented.

Large amounts of synthetic MC training data were generated to introduce variability across multiple dimensions like signal variability, signal location, SNR, and background environments. This variability in the training data improves robustness along those dimensions in the learned models. Evaluating on synthetic data generated on top of real recorded RF backgrounds, the system detected elaborate wideband signals with high accuracy, by capturing the large-scale spectro-temporal patterns of SOIs such as drones, Bluetooth and WiFi.

The proposed architecture was able to capture a range of RF activity across multiple temporal and frequency scales using a hierarchical processing approach and segmental scoring to produce robust detections, and offered a degree of robustness to false alarms in the presence of significant background RF activity. While we see many areas for further improvement, like dealing with lower SNRs., using a more sophisticated front end that use phase information, and achieving better narrowband detection performance, the proposed system was aimed to establish an initial framework for wideband SOI detection, that would be further improved in future work. In its current form it showed good performance for several types of elaborate signals and represents a promising approach to the wideband spectral monitoring problem using state-of-the-art deep learning approaches.

ACKNOWLEDGEMENTS

This material is based upon work supported by the Defense Advanced Research Projects Agency (DARPA) and Space and Naval Warfare Systems Center, Pacific (SSC Pacific), under Contract No. N66001-18-C-4044. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of DARPA or SSC Pacific. We thank Robert Oxsen for his help with the upsampling code. We also thank Paul Tilghman, Esko Jaska, and Joshua Alspector for many valuable suggestions.

REFERENCES

[1] I. Goodfellow, Y. Bengio, & A. Courville, *Deep Learning*. Cambridge, MA, USA: MIT Press, 2016.

[2] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, et al., ImageNet Large Scale Visual Recognition Challenge, *International Journal of Computer Vision*, vol. 115, pp. 211-252, 2015.

[3] L. Deng, J. Li, J.-T. Huang, K. Yao, D. Yu, F. Seide, M. Seltzer, G. Zweig, X. He, J. Williams, Y. Gong, & A. Acero, Recent Advances in Deep Learning for Speech Research at Microsoft, *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)* 2013.

[4] M. Kulin, T. Kazaz, I. Moerman, E.D. Poorter, End-to-end Learning from Spectrum Data: a Deep Learning Approach for Wireless Signal Identification in Spectrum Monitoring Applications, *IEEE Access* 6 (2018) 18484–18501.

[5] N.E. West & T. O'Shea, Deep Architectures for Modulation Recognition, *Proc. of IEEE International Symposium on Dynamic Spectrum Access Networks (DySPAN)*, 2017, pp. 1–6, Baltimore, MD, USA.

[6] T.J. O'Shea &, J. Corgan, Convolutional Radio Modulation Recognition Networks, 2016. arXiv: 1602.04105.

[7] DARPA: Radio Frequency Machine Learning Systems (RFMLS) https://www.darpa.mil/program/radio-frequency-machine-learning-systems

[8] K. He, G. Gkioxari, P. Dollar, & R. Girshick. Mask R-CNN. *IEEE International Conference on Computer Vision*, pp. 2980–2988, 2017.

[9] Mask R-CNN for Object Detection and Segmentation, https://github.com/matterport/Mask_RCNN.

[10] K. He, X. Zhang, S. Ren & J. Sun, Deep Residual Learning for Image Recognition, *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, 2016, pp. 770-778, doi: 10.1109/CVPR.2016.90.

[11] K. He, X. Zhang, S. Ren, & J. Sun, Identity Mappings in Deep Residual Networks, *European Conference on Computer Vision (ECCV)*, 2016.

[12] T. He, Z. Zhang, H. Zhang, Z. Zhang, J. Xie & M. Li, Bag of Tricks for Image Classification with Convolutional Neural Networks, *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Long Beach, CA, USA, 2019, pp. 558-567, doi: 10.1109/CVPR.2019.00065.

[13] V. Badrinarayanan, A. Kendall, & R. Cipolla, Segnet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation, *IEEE Transactions on Pattern Analysis and Machine Intelligence* 39.12 (2017), pp. 2481-2495.

[14] M. L. Seltzer, D. Yu and Y. Wang, "An investigation of deep neural networks for noise robust speech recognition," 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, Vancouver, BC, 2013, pp. 7398-7402, doi: 10.1109/ICASSP.2013.6639100.