Learning Robust General Radio Signal Detection using Computer Vision Methods

Tim O'Shea*, Tamohgna Roy*, T. Charles Clancy*
*Bradley Department of Electrical and Computer Engineering, Virginia Tech, Arlington, VA
{oshea,tamoghna,tcc}@vt.edu
†DeepSig Inc., Arlington, VA
{tim,charles}@deepsig.io

***Abstract*—We introduce a new method for radio signal detection and localization within the time-frequency spectrum based on the use of convolutional neural networks for bounding box regression. Recently, this class of approach has surpassed human-level performance on computer vision benchmarks for object detection, but similar techniques have not yet been adopted for radio applications. We introduce the basic approach explain how labeled training data containing wideband spectrum annotated with masks and bounding boxes can be used to train a highly effective radio signal detector which achieves higher levels of contextual understanding and improved sensitivity performance when compared with more traditional nave energy thresholding based signal detection schemes. We extend prior work from the computer vision field, employing a variation of the You Only Look Once (YOLO) architecture which is a fast and accurate visual object detector. Results are shown from illustrating the effectiveness from our entry into the DARPA Battle-of-the-ModRecs competition and over the air datasets.**

I. Introduction

Signal detection and localization are critical roles in many wireless systems. In radio communications and sensor systems these tasks defines the ability of a system to accurately build and maintain an up-to-date view of their current operating environment. This is foundational for detecting and interoperating with nearby users, detecting and isolating sources of interference, flagging significant spectral events, or identifying spectral vacancies within the radio spectrum. Unfortunately sensing systems can be difficult to deploy robustly in heterogeneous real-world wireless applications due largely to their over-specificity or lack of sensitivity as we discuss below. A new class of data-driven radio detectors, which leverage the powerful new techniques developed in computer vision hold the potential to greatly improve the practical performance and usefulness of such systems by embracing data-centric strategies for sensing algorithm design. In doing so, it may finally be possible to achieve sensing systems which generalize well, are resilient to impairments, and achieve good sensing performance in a wide range of scenarios.

II. Background

Algorithmic approaches to signal detection in radio communications systems have generally fallen into two classes: highly general methods which achieve relatively poor constant false alarm rate (CFAR) performance, and highly specialized methods which offer strong sensitivity, but are specific to only one type of signal, modulation property, or set of channel assumptions (low frequency offset, etc). An example of the first class is that of simple energy detection, which is commonly done by level thresholding on power spectral density frequency-bins. Such an approach can be easily cast into convenient probabilistic form for analysis, but is severely constrained in its ability to leverage any additional information about the context or structural patterns of emissions to improve discrimination. The second class focuses on methods such as matched filters, cylco-stationary energy detectors, and other methods which achieve high sensitivity but are specific to certain signals or signal properties they were designed for. Since both have favorable properties and differing complexity, complex schemes for dynamic switching between them based on further SNR estimation have been investigated [1]. Our approach seeks to introduce a third class of detectors which achieves both generality and sensitivity with a single set of algorithms, allowing for robust detection of many types of emissions without significant manual tuning, specializing detection and sensitivity for the signals present, but doing so in a way which is only constrained by available data.

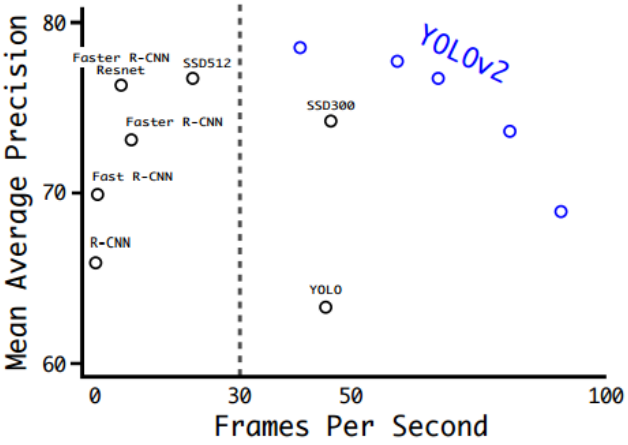

Fig. 1. Speed vs Accuracy for different DNN based object Detectors [8]

III. Technical Approach

Deep Neural Network (DNN) based approaches have achieved state of the art performances in computer vision

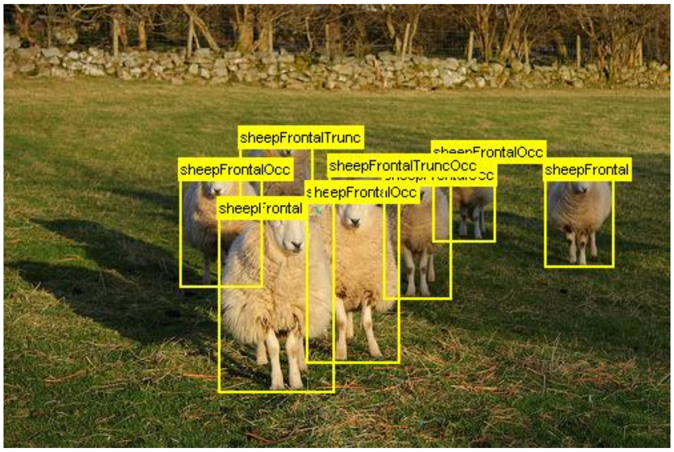

Fig. 2. Example bounding box annotations on sheep from VOC12 [11]

tasks such as image classification, object detection and object localization, outperforming humans on many of these tasks [3], [9]. Signal detection and localization are very analogous problems to visual object detection, and human visual performance in this task is already quite good when considering manual decisions made while looking at a spectrogram. A number of techniques exist for DNN based object detection, which have been steadily improving in recent years. Figure 1 compares the speed-accuracy trade-off for these various DNN based visual object detectors, where accuracy is measured on PASCAL VOC Dataset 2007 [10].

The Faster RCNN [4] based approach coupled with Deep Residual Networks [5] provides state of the art performance in object detection and localization tasks. This typically takes the form of predicting bounding boxes, labels and likelihoods of a collection of objects within an image such as the one shown in figure 2. Many of the approaches such as RCNN have an iterative component, running multiple network forwards passes to refine bounding box predictions over multiple objects in different regions. However, the superior performance of these approaches comes at the cost of significant computational complexity and reduction in classifier speed and throughput as can be seen in figure 1.

Detection speed often of prime importance for radio emission detection and classification in our problem. Thus, relatively faster object detection schemes with single forward pass structures such as You Only Look Once (YOLO) [7], [8] or Single Shot Multibox Detector (SSD) [6] are the most viable options to pursue. Since the modified YOLO architecture [8] provides performance improvement over SSD, the former is leveraged as a basis for this work.

A. *Our Tiny-Yolo Network Architecture*

We use a network architecture for this work which is a variant of YOLO (known as tiny-YOLO) as is described in I. Note that this network is much smaller than the full-size one used in [8]. Compared to visual object recognition tasks, recognition of spectral events is a relatively simpler task in many cases, allowing for smaller networks to be used. Additionally, a smaller network helps to reduce overfitting on the currently available smaller datasets for the task, and reduces the computational complexity of forwards passes, resulting in lower power and faster operation.

TABLE I. TABLE INPUT/OUTPUT SHAPES

Layer Number	Layer Type	Kernel Size	Number of Feature Maps
1,2,3,4,5,6	Conv+Maxpool	(3,3)	16,32,64,128,256,512
7,8	Conv	(3,3)	1024,1024
9	Conv	(1,1)	30

The network takes raw spectral power values in time and frequency as input, and produces an output grid of detections through a series of narrowing convolutional layers with increasing numbers of filters. Each grid cell in the output contains regression targets for relative height and width, and class likelihood, allowing the network to simultaneously detect many emissions within a single spectrogram, where the number is a function of the grid size and the number of possible detection per grid location. Anchor boxes may also be used to improve performance as shown in [8], however we do not include these in or initial implementation.

There are still a number of design parameters such as the spectrogram dimensions used, the time covered by a single spectrogram, the network architecture choice, and tuning of loss function parameters, all of which have not been fully optimized yet well for this task. Optimization of these design parameters will be key to obtaining the best performance for a specific application area.

B. *Training & Evaluation Dataset*

The direct training process for network weights relies completely on supervised learning, where a series of bounding boxes and class labels are annotated onto wide-band RF spectral data. We develop such a dataset of labeled wide-band radio data with bounding boxes in time-frequency on top of signals of interest, by building a simulator in GNU Radio. This same simulator was developed for and used by our to prepare for the DARPA Battle-of-the-ModRecs event. It employed 20+ single carrier modulation types spread randomly across a wide band with random emission times and intervals. The same approach can be conducted on real world training datasets but requires manual annotation of signal regions in order to bootstrap the training process. This approach proved extremely effective during the DARPA run competition, providing robust wide band burst detection which we believe outperformed traditional entries. Unfortunately, the DARPA contest infrastructure was deemed a failure, the results were never announced, and the dataset has not yet been released.

The training loss function is a bit involved, but generally involves the minimization of a sum of scaled mean-squared error loss between target labels and network predictions for each of the grid output values. The full loss function from [7] is shown below in equation 1 where $\mathscr{L}_{YOLO}$ represents the aggregate loss function optimized by the network, λ represents a scaling factor for each sub-objective, D_{L2} represents the L2 distance between values (x/y location and width/height (w/h)), C_i represents a class presence, and $p_i(c)$ is the class likelihood. However for most of our work we can simplify this expression to a single class (i.e. detection only, not classification).

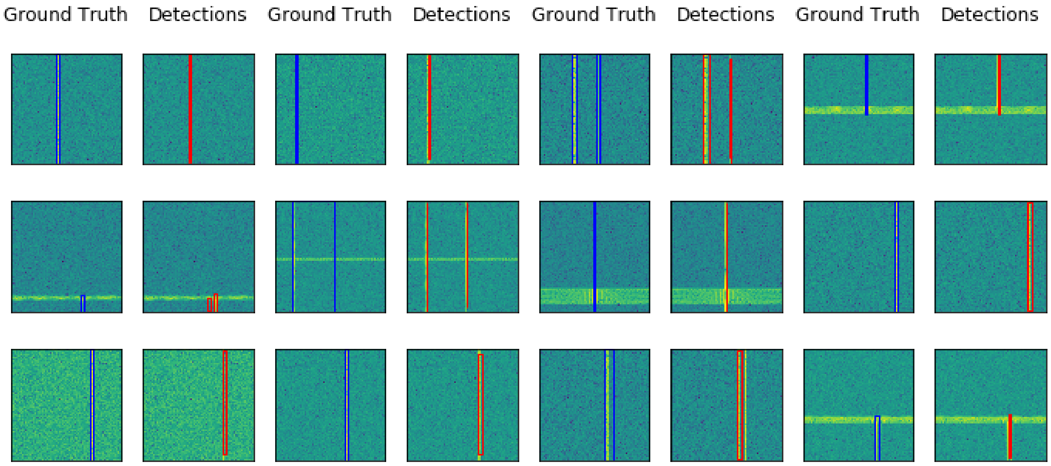

Fig. 3. Synthetic training set spectrograms with ground truth and predictions from the Tiny-YOLO network

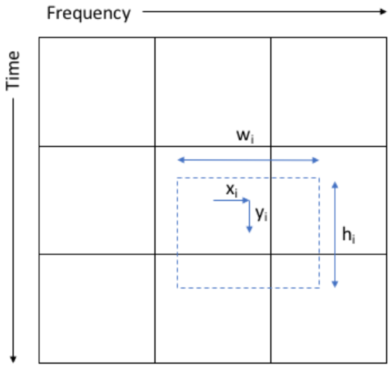

Fig. 4. Bounding box regression variable layout per grid-cell

$$
\begin{aligned}
\mathscr{L}_{YOLO} = \lambda_c \sum_{i=0}^{S^2} \sum_{j=0}^{B} \nVdash_{ij}^{obj} D_{L2}^2((x_i, y_i), (\hat{x}_i, \hat{y}_i)) \\
+\lambda_c \sum_{i=0}^{S^2} \sum_{j=0}^{B} \nVdash_{ij}^{obj} D_{L2}((w_i, h_i), (\hat{w}_i, \hat{h}_i)) \\
+\sum_{i=0}^{S^2} \sum_{j=0}^{B} \nVdash_{ij}^{obj} (C_i - \hat{C}_i)^2 \qquad (1) \\
+\lambda_{no-obj} \sum_{i=0}^{S^2} \sum_{j=0}^{B} \nVdash_{ij}^{no-obj} (C_i - \hat{C}_i)^2 \\
+\sum_{i=0}^{S^2} \sum_{c \in classes} (p_i(c) - \hat{p}_i(c))^2
\end{aligned}
$$

Using this loss function, standard network back-propagation techniques can be used. We use the Adam [2] stochastic gradient descent optimizer to update network weights until we've reached a stopping point.

IV. Performance Analysis

The Tiny-Yolo network described in the previous section was trained on 20,000 spectrograms, each of which contained at least one event. A batch size of 64 was chosen and the network was trained for 40,000 iterations. Figure 3 shows the ground truth images (with bounding box labels shown in blue) from the validation set along with the YOLO output (predicted bounding boxes shown in red).

The results show that the network is able to detect and localize the spectral events accurately. Note that the horizontal patches (wideband distortion/interference) are correctly not detected as spectral events by the network, where a naive energy based detector would likely have classified those regions as spectral events as well. As the output of such a network is a probability regression, we can apply traditional CFAR techniques to tuning false alarm rates and detection probabilities based on threshold levels directly on these outputs. Total detection time for 2000 spectrograms was 6 seconds on a single desktop machine with Intel i7-7700 processor and NVIDIA GTX 1080 GPU which translates to around 3 ms per spectrogram. In this way, such a system could immediately be used for real-time operation in a number of scenarios and sample rates, even on older generation GPU hardware using this approach.

V. Conclusion

Mature computer vision based detection methods such as this hold enormous promise in the radio sensing domain. Our initial investigation using synthetic data has produced robust predicted bounding box detections within complex dynamic wideband environments. Significant remaining work remains to be done in order to quantify the detection sensitivity, and to evaluate performance against baseline methods and annotated over the air radio data.

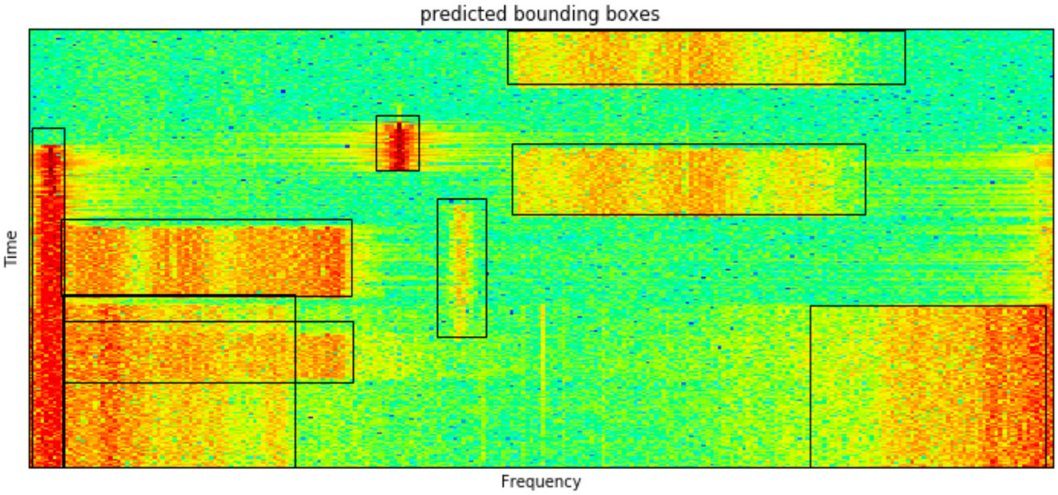

Fig. 5. Small-Yolo network predictions of signal bounding boxes on over the air ISM band radio emissions

Our evaluation of this model while training on annotated real world data sets containing RF recordings in the ISM band, shown in figure 4, yield promising initial performance as well. This example illustrates a number of features which are extremely appealing such as resilience to detection under interference, ability to delineate very close signals, and detection of complete signals with difficult impairments such as largely faded frequency sub-bands. Each of these are cases which would give simple energy detection methods trouble, resulting in incorrect merging or splitting of signal prediction bounds. This lends significant hope to the resilience and performance improvement this method holds for future sensing systems. There are numerous additional optimizations and training strategies which can help the performance of such a system, but as we are still working to perfect this approach at DeepSig, we leave an in depth exploration of these parameters for future work.

References

[1] H. Kim and K. G. Shin, "In-band spectrum sensing in cognitive radio networks: Energy detection or feature detection?" In *Proceedings of the 14th ACM international conference on Mobile computing and networking*, ACM, 2008, pp. 14–25.

[2] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[3] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1026–1034.

[4] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, Eds., pp. 91–99, 2015. [Online]. Available: http://papers.nips.cc/paper/5638-faster-r-cnn-towards-real-time-object-detection-with-region-proposal-networks.pdf.

[5] J. Dai, Y. Li, K. He, and J. Sun, "R-fcn: Object detection via region-based fully convolutional networks," in *Neural Information Processing Systems (NIPS), 2016*, 2016.

[6] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "Ssd: Single shot multibox detector," in *Computer Vision – ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I*, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds. Cham: Springer International Publishing, 2016, pp. 21–37, ISBN: 978-3-319-46448-0. DOI: 10.1007/978-3-319-46448-0_2. [Online]. Available: http://dx.doi.org/10.1007/978-3-319-46448-0_2.

[7] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 779–788.

[8] J. Redmon and A. Farhadi, "Yolo9000: Better, faster, stronger," *arXiv preprint arXiv:1612.08242*, 2016.

[9] R. Ewerth, M. Springstein, L. A. Phan-Vogtmann, and J. Schütze, "are machines better than humans in image tagging?i-a user study adds to the puzzle," in *European Conference on Information Retrieval*, Springer, 2017, pp. 186–198.

[10] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, *The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results*, http://www.pascal-network.org/challenges/VOC/voc2007/workshop/index.html.

[11] ——, *The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results*, http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html.