Transparency in Automatic Modulation Recognition Applications with GradCAM-based Analysis

Ozan Gülbaş
Email: ozan.gulbas@student.umons.ac.be

Department of Electromagnetism and Telecommunication
Mons Polytechnic Faculty
Mons, Belgium

***Abstract*—AMR is the act of specifying the modulation type of a detected Radio Frequency (RF) signal at a given time, space, and frequency. This kind of application plays a vital role in handling the dynamic spectrum and preventing the interference besides usage of an Artificial Intelligence (AI) to achieve fast and efficient classifications, is rather common. In this paper, the usage of XAI to help create a much more trustworthy environment in such applications is considered. A Convolutional Neural Network (CNN) model is used to achieve robustness in the computation of the classification of In-Phase/Quadrature (I/Q) dataset samples. The proposed model is highly convenient to use since CNNs are capable to perform like state-of-the-art models in the computation of Time Series datasets, which are quite similar to I/Q datasets, and their structural ease of implementation advantage for Class Activation Mapping (CAM). CAM is used for calculating which features of the AI model have most contributed to the result and displaying them in a visually understandable way to make predictions easily interpreted by humans.**

***Index Terms*—XAI, AMR, AMC, Grad-CAM, Keras, Tensorflow, Oshea**

I. Introduction

ARTIFICIAL intelligence (AI) has revolutionized various fields, from computer science to gastronomy, social sciences to politics, and engineering. In the domain of communications engineering, automatic modulation recognition (AMR), which is one of the fundamental topics discussed in this paper, has been highly influenced and evolved by AI. With the increasing need for fast, accurate, and efficient classification in AMR applications, AI-based solutions are more in demand. However, the interpretability of AI-based solutions is mostly questioned because of the lack of transparency.

In this paper, we address these concerns and try to find a solution by employing GradCAM-based analysis, a method that offers visual explanations for the choices made by AI model [1], which is an Explainable Artificial Intelligence (XAI) technique to enhance interpretability in AI applications. With the constructed CNN, getting high accuracy in classifying I/Q dataset samples is aimed.

The ability of CNNs to precisely analyze time series data, which is very similar to I/Q data used in AMR applications, is very well known. [2] Furthermore, the architectural simplicity of CNNs made possible the implementation of Class Activation Mapping (CAM), which is the backbone of GradCAM. With the use of CAM, identifying the features that contribute the most to the AI model's prediction can be done. Moreover, obtained CAM results can be used to visualize the predictions made by AI in a human-understandable way.

With the GradCAM-based analysis, improving the modulation recognition accuracy for AMR applications and providing interpretability is expected. By displaying visual cues that highlight the important features, we seek to close the gap between AI's prediction process and human comprehension.

II. Convolutional Neural Networks

CNNs are becoming to have high accuracy performance on time series data as well as their known performance on image data. In computer vision applications such as semantic labeling, object identification, and image segmentation, CNNs have demonstrated outstanding performance, thanks to their ability to capture spatial relations and preserve spatial information throughout processing. Compared to ResNet, which is a recurrent neural network (RNN) that performs exceptionally well with the time series data, its capability to reach that kind of performance and being easier to train with the time series data [2], [4], and structural compatibility to class activation mapping (CAM) applications are main aspects to use of CNN for the application described in this paper.

CAM is a method in AI terminology used for locating discriminative regions that influenced AI on its prediction, through the given input. Due to CNNs' built architectural qualities, which allow them to learn spatial representations and generate precise CAM outputs, they are highly compatible with the explainability applications in time series and I/Q data samples.

The used CNN in this paper consists of two convolutional layers. The two convolutional layers have 256 and 80 filters and window sizes of (1, 3) each. No padding is applied on convolutional layers since the same output size as the input is desired. Pooling is not applied and the model has two dense layers. The final output shape is acquired with a softmax layer. A visual representation of the model is given in Figure 1.

This kind of structure becomes highly useful when it comes to CAM applications. The output of the last convolutional layer enables us to extract useful spatial information. The extracted information then can be used to calculate the dot product with the weights of the output layer, which is the

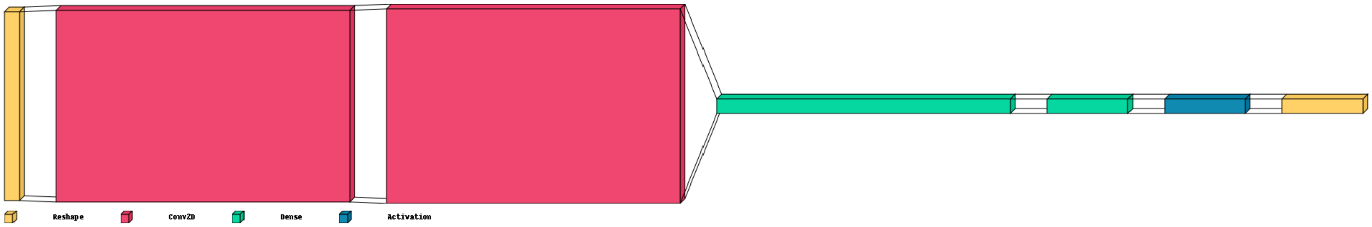

Fig. 1. The illustration of the proposed convolutional network generated with ”Visualkeras”[3]

CAM operation itself. The obtained result can be used to highlight the key areas of the given data in a visual, and have an improved comprehension of the input data.

III. Class Activation Mapping

CAM is a method used to implement interpretability in AI models. The most popular use case of CAM is generating heatmaps on input images that are fed through the image recognition CNNs. By looking at those heatmaps generated with CAM, we can learn more about the discriminative regions of an image that contribute most to the final prediction.

The outputs of the last layer, which are the feature maps, contain different patterns that the network has learned to recognize in the input images. The most contributing regions in the decision-making of the AI can be revealed by calculating the dot product between those feature maps and the corresponding weights of the output layer which is the CAM itself. Then this obtained result is upsampled to match the size of the input image to make the calculated CAM understandable. An easy-to-interpret and visually appealing heatmap that shows the areas which influenced the final prediction the most can be generated by superimposing the upsampled CAM onto the input image.

Although CAM is mostly associated with image recognition applications, the concept itself can also be used for analyzing other forms of data, such as time series and I/Q data. The principle is the same on a fundamental basis and does not depend on the data type.

Let $F_k(t)$ denote activation which is generated by the k^{th} filter at the timestep t in the last convolution layer and $w_{c,k}$ to represent the weight from the output of the final softmax layer, which is also generated by the k^{th} filter and the class c. We can write the input of the final softmax function as [2], [4]

$$g_c = \sum_k w_{c,k} \sum_t F_k\left(t\right) \tag{1}$$

We can write a class activation map for class c at a timestep t as

$$A_{c,t} = \sum_k w_{c,k} F_k\left(t\right) \tag{2}$$

Equation 2 further demonstrates how the architecture is especially suited for CAM application because of the zero-padding in the CNN. The individual influence score for each time step may be calculated since the output of the last convolutional block is the same length as the input [4].

IV. Automatic Modulation Recognition/Classification

A fundamental component of signal processing, automatic modulation recognition (AMR), also known as automatic modulation classification (AMC), plays its role between signal detection and demodulation. AMR is highly important in both military and civilian applications.

In the scope of military applications, AMR allows for the categorization and detection of modulation used in military communication systems, which is highly beneficial for various defense operations, such as electronic warfare, spectrum monitoring, and intelligence gathering.[5] In the study done by Iglesias, Grajal, and Yeste-Ojeda (2011)[6], the importance of AMR applications is shown with an AMR based on low complexity signal characteristics, designed for broadband military applications, to improve latency and the proportion of real-time operation .

AMR also has a wide range of use cases in wireless communication systems for civil purposes, and one of the most important use cases in the domain of civilian applications is the application of link adaptation (LA) [5]. The concept of automatically identifying the modulation schemes is becoming increasingly important with the spread of wireless technologies.

The process of AMR involves investigating the properties of the received signal, removing related characteristics, and using machine learning (ML) methods to categorize the received signal. This process is possible because the various modulation schemes, such as frequency modulation (FM), amplitude modulation (AM), phase shift keying (PSK), and quadrature amplitude modulation (QAM), have distinctive characteristics that can be analyzed with high precision.

Continuing research and developments in AMR techniques helped to develop intelligent communication systems, software-defined radios (SDR), etc. With the AMR, the process benefits from improved system performance, better spectrum usage, and more adaptability to changing communication contexts

To sum up this chapter, AMR has a critical role in communication system optimization, defense operations, and wireless network management due to its use cases in both military and civilian areas. As technology develops further, state-of-the-art communication systems leaning more and more on AMR techniques to enable more effective and reliable data transfer.

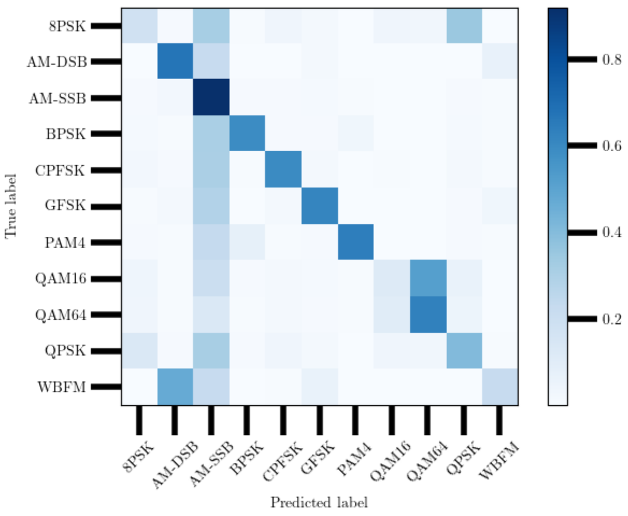

Fig. 2. Confussion matrix of the proposed model

V. Experiments and Discussion

The experiments done in this study were conducted with the previously proposed CNN model. The model was trained on the 'DeepSig Dataset: RadioML.10A,'[7] employing a 50/50 split for training and test/validation data. The training process lasted for 69 epochs, achieving an accuracy of %50. In the experiment, we focused on generating heatmaps from the test partition of the previously mentioned dataset. In particular, 60 heatmaps are generated using four different modulation schemes, including Quadrature Phase Shift Keying (QPSK), 8-Phase Shift Keying (8PSK), Continuous Phase Frequency Shift Keying (CPFSK), and Pulse Amplitude Modulation 4 (PAM4). Three samples were chosen for each of the five Signal-to-Noise Ratio (SNR) values—-18 dB, -10 dB, 0 dB, 10 dB, and 18 dB—within each modulation scheme. These created heatmaps help grasp a better understanding of the CNN model's decision-making process.

A. Observations on QPSK and 8PSK samples

When the QPSK and 8PSK samples were examined, one apparent finding is how the low SNR levels affect the produced heatmaps. It is clear that when the SNR value gets low, the heatmap's visual representation gets cluttered and presents several locations as relevant. This finding is not unexpected and is consistent with the researches conducted by Liu, Yang, and Gamal [8], and Tekbıyık, Ekti, Görçin, Kurt, Keçeci [9] using a variety of CNNs and datasets. They were not implementing any XAI method however, they showed the similar results on SNR and accuracy. Furthermore, these findings also point out the connection between accuracy and SNR. The resulting visual representations of heatmaps get increasingly distinct and clear as the SNR value increases. These findings are compatible with the overall finding of the experiments and further emphasizes the influence of SNR on interpretability in AMR applications.

Another important finding is the presence of sudden peaks and valleys in the QPSK and 8PSK data as contributing regions. This occurrence can be also interpreted as a pattern where the I/Q signal's highest peaks and lowest valleys tend to be displayed as more influential than other peaks and valleys present in the signal. Moreover, the confusion matrix (see Figure 2) is also shown that the model often has trouble with correctly distinguishing between QPSK and 8PSK data. The similarity can be seen between the two modulation methods' resulting heatmap visuals (see Figure 3).

B. Observations on CPFSK data

The accuracy of CPFSK data is higher than both QPSK and 8PSK, as shown in the confusion matrix (See Figure 2). As a result, the generated heatmap visuals resulted in greater clarity when compared with QPSK and 8PSK. However, there is a noticeable performance drop at 0 dB SNR that causes confusing heatmap visuals. The increase in the classification accuracy with CPFSK data is further supported by the vivid heatmap graphics also which shows that the CNN model is better able to distinguish CPFSK modulation from QPSK and 8PSK.

Unlike QPSK and 8PSK, where the highest peaks and lowest valleys are more influential, the influence of these extreme values is noticeably decreased in CPFSK. The final prediction is instead more strongly affected by the other aspects CPFSK has. Furthermore, when the given I/Q signal contains repeated patterns, the corresponding regions become more significant, showing the importance of these recurring patterns (see Figure 4). These findings draw attention to the particular features of CPFSK and may be used in future studies to further increase the accuracy of CPFSK recognition.

C. Observations on PAM4 data

Among the modulations observed in the experiments, the PAM4 data shows the highest accuracy. Furthermore, the SNR and the I samples of the PAM4 data, show an intriguing connection. As the SNR value increases, the regions which strongly influence the classification decision in the I samples become less noticeable (see Figure 5).

Moreover, in cases where a repeating pattern is present, similar to the observations made for CPFSK, the same contributing regions are displayed consistently in the final heatmap visualization. This occurrence can be interpreted as the demonstration of the accuracy and consistency of the model's identification of contributing regions.

Another unique pattern that has been encountered is when the taken sample has mirrored the I and Q samples. The resulting heatmap visual in these situations shows the same contributing regions with nearly the same relevance scores (see Figure 6). This consistency that is seen in the visual representations of the heatmaps also supports the model's capacity to recognize crucial patterns which can be used for improving the explainability of AMR applications in future studies.

VI. Conclusion

In this study, the use of XAI, specifically the Grad-CAM algorithm, in AMR applications is investigated. First, a CNN model is presented to be used in experiments, which is selected for its ease of trainability and compatibility in the implementation of CAM applications. Next, CAM and AMR are briefly explained, addressing ”what it is?” and ”how it works?” aspects. Then, some observations are performed based on the results of the experiments. Remarkably, experiments showed even with a limited dataset, apparent patterns, and repeating behaviors can be found, which emphasizes the potential of the presented method to improve interpretability in AMR applications. Concluded research provides insightful information about the mechanisms underlying modulation recognition applications performed by AI, through the use of Grad-CAM for enhanced comprehension in this field.

Acknowledgement

I would like to thank Dr. Ir. Professor Véronique Moeyaert, Ph.D. student Ir. Alexander Gros and Masters student Florian Facchin for the help provided during the realization of this project.

References

[1] D. Reiff. “Understand your algorithm with grad-CAM,” Medium. (May 12, 2022), [Online]. Available: https://towardsdatascience.com/understand-your-algorithm-with-grad-cam-d3b62fce353 (visited on 05/25/2023).

[2] Z. Wang, W. Yan, and T. Oates, *Time series classification from scratch with deep neural networks: A strong baseline*, Dec. 14, 2016. arXiv: 1611.06455[cs,stat]. [Online]. Available: http://arxiv.org/abs/1611.06455 (visited on 05/25/2023).

[3] P. Gavrikov, *Visualkeras for keras / TensorFlow*, original-date: 2020-09-25T10:11:42Z, May 25, 2023. [Online]. Available: https://github.com/paulgavrikov/visualkeras (visited on 05/26/2023).

[4] K. Wickstrøm, K. Ø. Mikalsen, M. Kampffmeyer, A. Revhaug, and R. Jenssen, “Uncertainty-aware deep ensembles for reliable and explainable predictions of clinical time series,” *IEEE Journal of Biomedical and Health Informatics*, vol. 25, no. 7, pp. 2435–2444, Jul. 2021, ISSN: 2168-2194, 2168-2208. DOI: 10.1109/JBHI.2020.3042637. arXiv: 2010.11310[cs,eess]. [Online]. Available: http://arxiv.org/abs/2010.11310 (visited on 05/25/2023).

[5] Z. Zhu, “Automatic modulation classification,”

[6] V. Iglesias, J. Grajal, and O. Yeste-Ojeda, “Automatic modulation classifier for military applications,”

[7] D. Inc. “RF datasets for machine learning — DeepSig.” (), [Online]. Available: https://www.deepsig.ai/datasets (visited on 05/25/2023).

[8] X. Liu, D. Yang, and A. E. Gamal, “Deep neural network architectures for modulation classification,” in *2017 51st Asilomar Conference on Signals, Systems, and Computers*, Pacific Grove, CA, USA: IEEE, Oct. 2017, pp. 915–919, ISBN: 978-1-5386-1823-3. DOI: 10.1109/ACSSC.2017.8335483. [Online]. Available: http://ieeexplore.ieee.org/document/8335483/ (visited on 05/25/2023).

[9] K. Tekbıyık, A. R. Ekti, A. Görçin, G. K. Kurt, and C. Keçeci, “Robust and fast automatic modulation classification with CNN under multipath fading channels,” in *2020 IEEE 91st Vehicular Technology Conference (VTC2020-Spring)*, May 2020, pp. 1–6. DOI: 10.1109/VTC2020-Spring48590.2020.9128408. arXiv: 1911.04970[cs,eess, stat]. [Online]. Available: http://arxiv.org/abs/1911.04970 (visited on 05/25/2023).

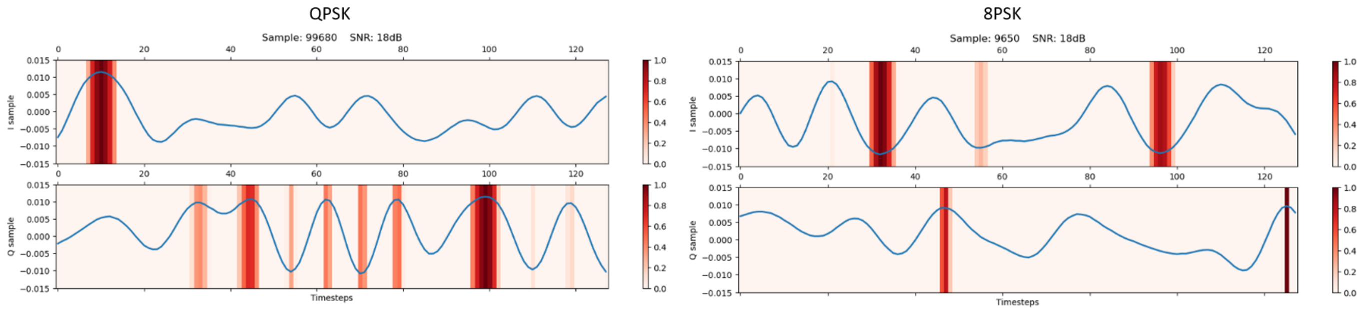

Fig. 3. QPSK and 8PSK comparison with the SNR value 18 dB.

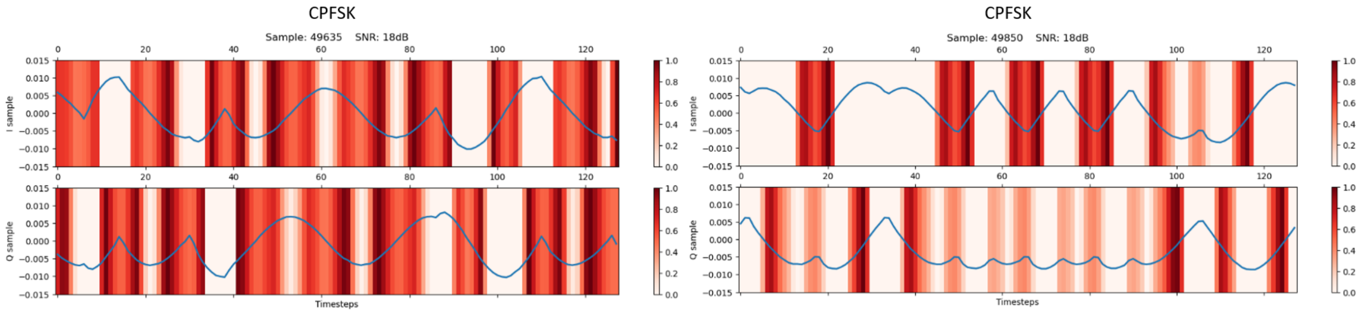

Fig. 4. Two samples of CPFSK that have repeating patterns with the SNR value 18 dB.

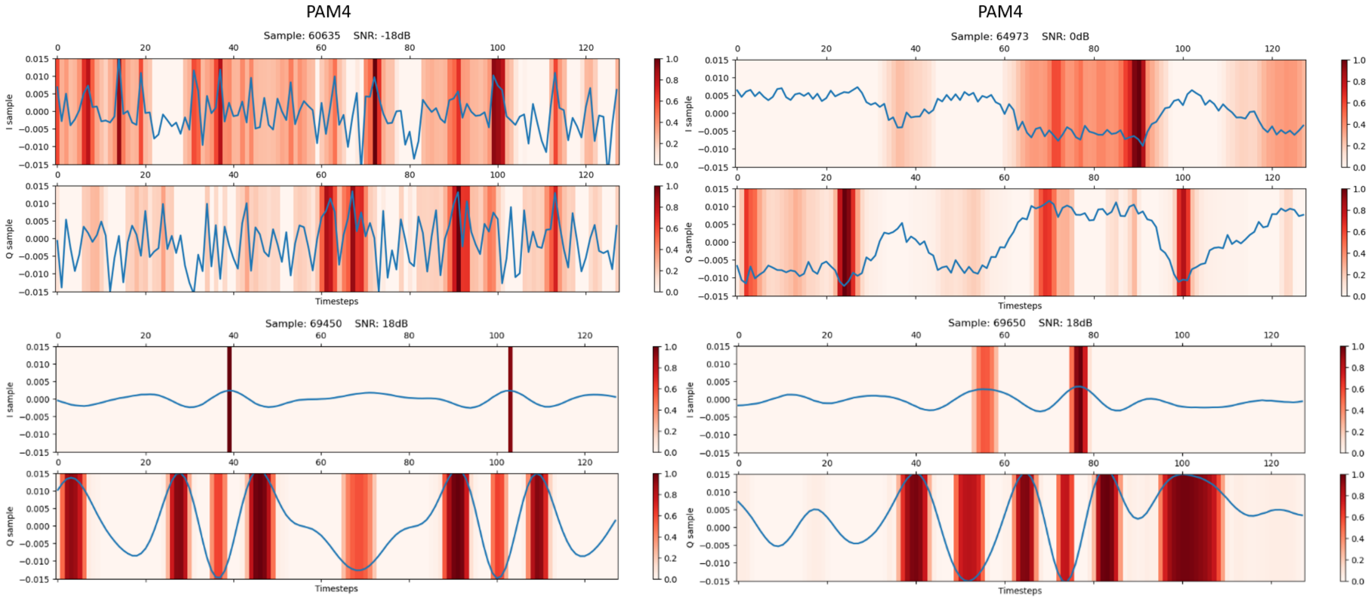

Fig. 5. Four samples of PAM4 with the decreasing SNR to show the relation between the I sample and the SNR.

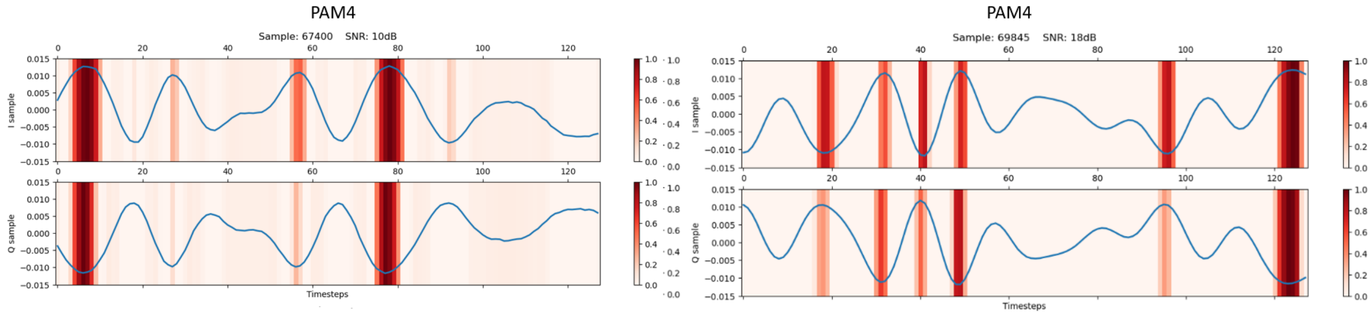

Fig. 6. Two samples of PAM4 that have mirrored I and Q samples.