Analytical Interpretation of the Gap of CNN's Cognition between SAR and Optical Target Recognition

Zhenpeng Feng[a], Hongbing Ji[a,*], Miloš Daković[b], Mingzhe Zhu[a] and Ljubiša Stanković[b]

[a]School of Electronic Engineering, Xidian University, Xi'an, China
[b]Faculty of Electrical Engineering, University of Montenegro, Podgorica, Montenegro

ARTICLE INFO

Keywords:
keywords-1 SAR imaging
keywords-2 target recognition
keywords-3 interpretable CNN
keywords-4 multi-order interaction

ABSTRACT

Synthetic aperture radar (SAR) automatic target recognition (ATR) is a crucial technique utilized in various scenarios of geoscience and remote sensing. Despite the remarkable success of convolutional neural networks (CNNs) in optical vision tasks, the application of CNNs in SAR ATR is still a challenging area due to the significant differences in the imaging mechanisms of SAR and optical images. This paper analytically addresses the cognitive gap of CNNs between optical and SAR images by leveraging multi-order interactions to measure their representation capacity. Furthermore, we propose a subjective evaluation strategy to compare human interactions with those of CNNs. Our findings reveal that CNNs operate differently for optical and SAR images. Specifically, for SAR images, CNNs' representation capacity is comparable to that of humans, as they can encode intermediate interactions better than simple and complex ones. In contrast, for optical images, CNNs excel at encoding simple and complex interactions, but not intermediate interactions.

1. Introduction

Synthetic aperture radar (SAR) imaging technology has been extensively utilized in the field of geoscience and remote sensing [2, 5] due to its remarkable capability to generate high-resolution images, even under conditions of low visibility such as dark nights or precipitation/foggy weather [21, 6]. Automatic target recognition (ATR) in SAR images has gained considerable attention owing to its wide-ranging applications in both military and civilian fields [14]. Like optical image processing, SAR ATR typically consists of three sequential stages: detection [16], discrimination [13], and classification [22]. During the detection stage, targets and other objects such as trees, buildings, and streetlights are identified. Subsequently, the discrimination stage utilizes various properties such as texture, size, and contrast to differentiate between the potential target pixels and other objects to determine if they originate from a genuine target. Finally, in the classification stage, the potential target is categorized into its most probable class.

The remarkable success of deep neural networks, particularly convolutional neural networks (CNNs), in computer vision has led to their application in SAR ATR, particularly in the detection and classification stages [8]. However, the internal workings of CNNs remain obscure, creating a gap between human cognition and the CNN's comprehension of SAR images [10]. Consequently, this limitation may impede the applicability of CNNs in certain life/health-relevant scenarios, such as wild rescue and military surveys [20, 18]. To address this issue, several CNN-interpretation algorithms have been developed to produce heatmaps that visualize the mechanisms of CNN's hidden layers, including class activation mapping (CAM) [25, 4], layer-wise relevance propagation (LRP) [1], RISE [15], LIME [17], etc. Specifically, these heatmaps can identify the importance of parts of an image for the CNN's decision. Nevertheless, these interpretations have two limitations: (1) they only provide qualitative interpretation instead of quantitative metrics and (2) they can only visualize the representations modeled by CNN, without clarifying which type of information is more suitable or unsuitable for CNN to model. In this case, these limitations make it difficult to entirely trust the interpretation results, and it can be challenging to use them to guide the design and training of CNNs. To alleviate these two limitations, multi-order interaction is proposed by Q. Zhang et al. to quantify the interaction utility between variable pairs on optical images [3]. They demonstrates that CNN is more likely to encode both too complex and too simple interactions, instead of encoding interactions of intermediate complexity, which is quite different from human's cognition. Nevertheless, this conclusion is not completely applicable to SAR images due to the peculiar characteristics of SAR images compared with optical images. In this paper, we utilize multi-order interaction to explain CNN's mechanism in classification stage of SAR ATR. Our findings reveal that for SAR images, CNNs are better in encoding intermediate interactions rather than simple and complex interactions (as shown in Fig. 1), which is different from optical images. The main contributions of this paper can be summarized as:

- It is the first attempt to utilize multi-order interaction to provide an analytical interpretation of CNNs' mechanism in SAR ATR.
- We reveal the difference of CNN's cognition between SAR images and optical images, i.e., CNN shows very similar learning and classifying mechanism to human provided SAR images but completely opposite mechanism provided optical images.

*Corresponding author: Hongbing Ji
hbji@xidian.edu.cn (Hongbing Ji)
ORCID(s):

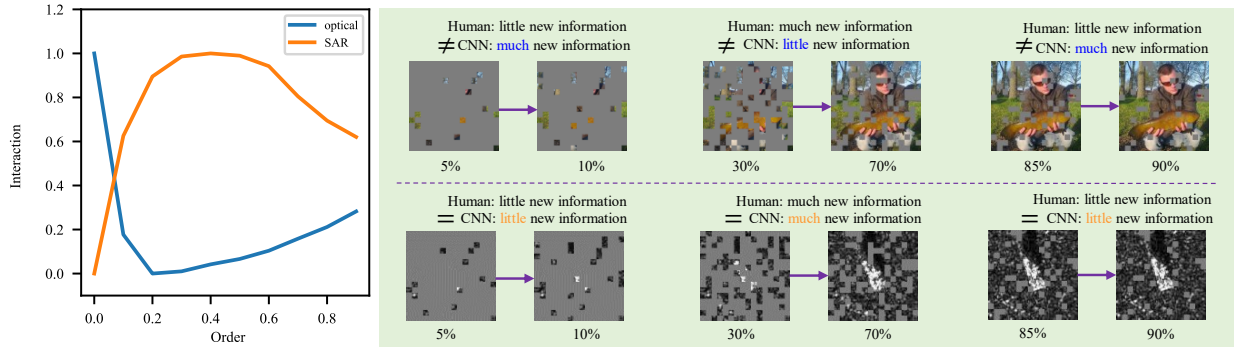

Figure 1: The comparison of CNN's representation bottleneck between optical images and SAR images. The multi-order interaction of optical images and SAR images (left). The cognition gap between CNN and humans in optical images and SAR images, respectively (right).

2. Representation Bottleneck of CNN

First, let's consider two classic questions about interpreting CNNs' mechanism in optical computer vision tasks:

- Are there any common tendencies of CNNs in representing specific types of features?
- Does a CNN encode similar semantic concepts as human beings for image classification?

To address the above two questions,we investigate the bottleneck of feature representation in CNNs, i.e., which types of concepts are likely to be encoded by a CNN, and which types of concepts are difficult to be learned. Interaction between input variables is usually regarded as an effective tool to analyze the feature representation of CNN since different variables are commonly combined to affect the CNN's inference (e.g., the inference of a vehicle can be explained as the interactions between left and right wheels, between body and barrel, etc.). Interactions can be understood as follows. Take an example of tiger, we let ϕ_{ears} measure the importance of the ears region i to the classification score of "tiger" class. Then the interaction between the ears region i and eyes region j is measured as the change of ϕ_{ears} by the absence or presence of the eyes region j. If the presence of j increases the importance ϕ_{ears} by 0.2, then, we consider 0.2 as the interaction value between regions i and j.

In recent decades, the investigation of interactions between input variables of neural networks has gained widespread attention. One approach for understanding these interactions is through building tree ensemble explanations for deep neural networks (DNNs), as demonstrated by Lundberg et al. [12]. Other methods include the extension of Integrated Gradients by Janizek et al. [9], which focuses on explaining pairwise feature interactions in DNNs. Additionally, Q. Zhang et al. [3] proposed the concept of multi-order interaction, which was initially used to understand and improve dropout. Subsequently, Q. Zhang et al. utilized game-theoretic interactions to develop a theoretical system for explaining the representation capacity of a DNN. This system includes explaining the generalization ability, adversarial transferability, and adversarial attacks of a DNN, as well as explaining the concepts encoded in a DNN [23, 24, 3]. In this paper, we leverage the algorithm proposed in [3] as an analysis tool to investigate the representation bottleneck in convolutional neural networks (CNNs) for both optical and SAR images.

3. Methodology

3.1. Cognition Mechanism of CNN

Now we turn to the scenario of SAR ATR, the above two questions will be further specified to:

- What is the common tendency of CNNs in representing specific types of features from SAR images?
- Does a CNN encode similar visual concepts as optical image classification?

To answer these questions, we utilize interaction to study the representation bottleneck of CNNs in SAR ATR. Given a pretrained CNN f, and an input SAR image with n pixels/patches $\mathbb{N} = \{1, 2, 3, \cdots, n\}$, $f(\mathbb{N})$ denotes the CNN's output of all input pixels. The m-th order multi-order interaction between two input variables i and j, $I^{(m)}(i, j)$, $i, j \in N, i \neq j, 0 \leq m \leq n-2$, is defined as follows:

$$I^{(m)}(i,j) = \mathbb{E}_{\mathbb{S}\subseteq N\setminus\{i,j\},|\mathbb{S}|=m}[\Delta(\{i,j\},\mathbb{S})] \tag{1}$$

where $\Delta(\{i,j\},\mathbb{S}) = [f(\{i,j\}\cup\mathbb{S}) - f(\{i\}\cup\mathbb{S}) - f(\{j\}\cup\mathbb{S}) + f(\mathbb{S})]$. Here, $f(\mathbb{S})$ is the output score when we keep variables in $\mathbb{S}\subseteq\mathbb{N}$unchanged but replace variables in $N\setminus\mathbb{S}$ by the baseline value. value. (1) tells the m-th order interaction $I^{(m)}(i,j)$ measures the average interaction utility between variables i, j under all possible contexts consisting of m variables. Specifically, $\Delta(\{i,j\},\mathbb{S}) = [f(\{i,j\}\cup\mathbb{S}) - f(\{i\}\cup\mathbb{S}) - f(\{j\}\cup\mathbb{S}) + f(\mathbb{S})]$ quantifies the marginal effects (the importance) of the variable j that are changed by the presence or absence of the variable i. It represents the utilities of the collaboration between i, j in a context $\mathbb{S}$.

Because $I^{(m)}(i,j)$ measures the interaction between variables i and j encoded in CNNs with m contextual variables,

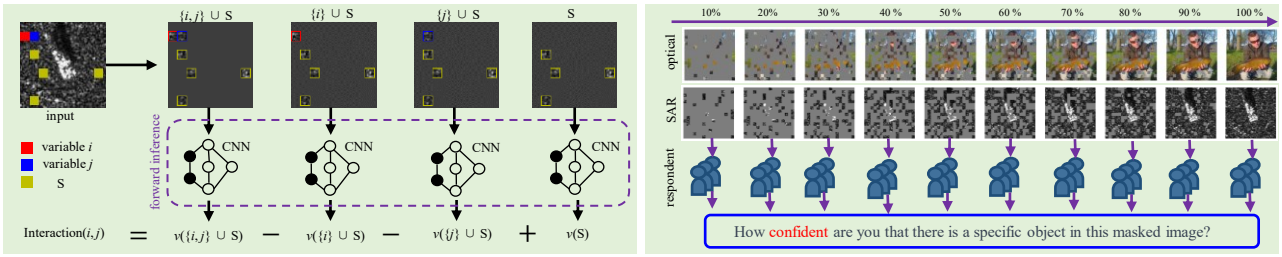

Figure 2: The illustration of multi-order interaction computation when the order $m = 4$ (top). The illustration of human's confidence differential (bottom).

we can consider the interaction utility $I^{(m)}(i, j)$ as a specific reason for the inference, which makes a compositional contribution to the output. In this way, a useful categorization of the underlying reasons for the network output can be achieved by classifying them based on their complexity. Specifically, simple interactions, which rely on only a few variables, can be categorized as simple underlying reasons and correspond to low-order interactions (m). Conversely, complex interactions, which depend on a large number of variables, can be considered as complex underlying reasons and correspond to high-order interactions (m). To assess the reasoning complexity of a deep neural network, we use a measure of the relative interaction strength $J(m)$ for the encoded m-th order interaction. This measure is defined as follows:

$$J(m) = \frac{\mathbb{E}_{x\in\Omega}[\mathbb{E}_{i,j}[I^{(m)}(i,j|x)]]}{\mathbb{E}_{m'}[\mathbb{E}_{x\in\Omega}[\mathbb{E}_{i,j}[I^{(m')}(i,j|x)]]]} \tag{2}$$

where Ω refers to the set of all samples, $I^{(m)}(i, j|x)$ denotes the interaction of variables i and j in a specific image $x \in \Omega$. $J(m)$ is normalized by the average value of interaction over all pairs of input variables in all samples. The distribution of $J(m)$ reflects the distribution of the complexity of interactions encoded in CNNs.

3.2. Cognition Mechanism of Human

Computing the interaction utility of human beings poses a significant challenge due to the complexity of the cerebral cortex's mechanisms and the lack of numerical metrics like multi-order interaction to measure the outputs. Therefore, we propose a subjective evaluation strategy as a means of investigating the representation bottleneck in the human brain, serving as a point of comparison. To implement this strategy, we randomly mask various ratios of pixels in images and present them to multiple respondents. The respondents are then asked to rate their confidence in the presence of a specific object in the masked image. The average score across respondents can be considered the confidence score for human classification if the information was unmasked in the image. Finally, we compute the first-order differential of the confidence scores across different masking ratios. This confidence differential provides insights into human brain's preference for the complexity of interaction encoding. A large value of confidence differential indicates the introduction of new information, while a small value implies that the masking had little effect on classification confidence.

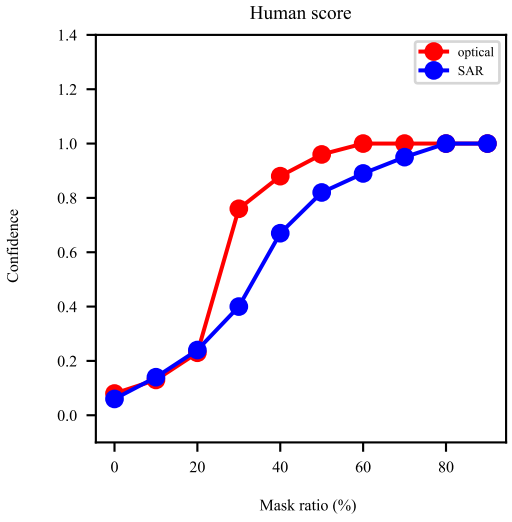

Figure 3: The human beings confidence score with different mask ratios for optical and SAR images.

4. Experiments

4.1. Implement

Dataset: In this paper, we evaluate our proposed approach on two well-established benchmarks: ILSVRC-1K and MSTAR. ILSVRC-1K is a widely-used benchmark in computer vision that comprises 1000 categories and 1.2 million optical images in the training set, and 50,000 in the validation set. MSTAR includes 2.5 thousand SAR images of 10 categories of ground vehicles in the training set and 2.6 thousand in the validation set. To ensure a fair comparison between the two benchmarks, we randomly selected 10 categories from the ILSVRC validation set as our optical images.

Details: In our experiments, the original dimension of input varibles is 224×224, i.e. $n = 224 \times 224 = 50176$. It is computationally infeasible to compute $J(m)$ by averaging all possible contexts $S \in N$, all pairs of variables $i, j \in N$, and all samples $S \subseteq N$. To address this issue, we adopt a simplification strategy inspired by previous work [3] to approximate $J(m)$. Specifically, we randomly select five im-

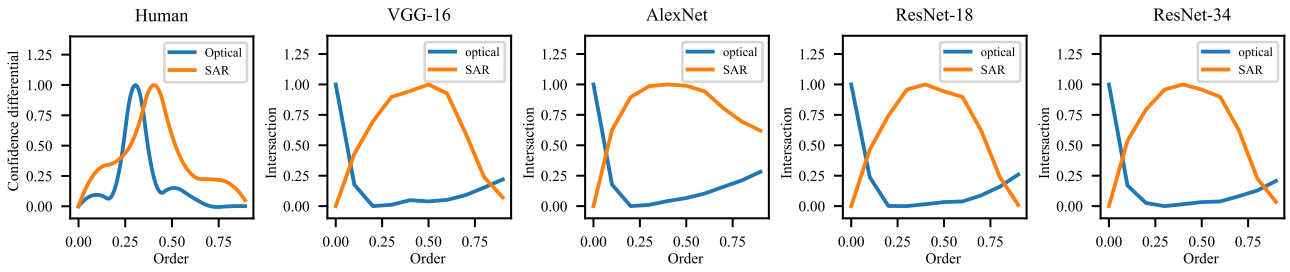

Figure 4: The differential of human's confidence score (first). The interaction curve of of AlexNet (second), VGG-16 (third), ResNet-18 (fourth) and ResNet-34 (fifth).

ages from each class and split each input image into 9×9 patches, resulting in $n = 81$ input variables. We impose the constraint that patch i must be located at the neighborhood patch j with a radius of one patch, without overlapping. It is because CNNs typically encode stronger interactions between neighbor patches. However, even with this simplification, computing $J(m)$ for all pairs of patches and all orders m is still computationally prohibitive, as the number of m collocations from 81 samples is also large. Therefore, we randomly sample 50 contexts S for each pair of patches i, j and each order m, where $|S| = m$. With this approach, the computational cost is manageable, requiring $9 \times 8 \times 50 = 3600$ computations of $I^{(m)(i,j)}$ for one image (approximately 9 minutes). We evaluate $J(m)$ for values of m ranging uniformly from 0 to $0.9n$ with step of $0.1n$.

In human's interaction evalution, each image is masked by the ratio of 10%, 20%, 30%, 40%, 50%, 60%, 70%, 80%, and 90%, as shown in Fig. 2 (right). We invited 20 individual respondents and they are required to give a score from 0 to 10 to represent how confident they are there is a specific object in the shown image. Here a larger score means higher confidence. The confidence scores are shown in Fig. 3.

4.2. Analysis of Results

We computed the relative interaction strength $J(m)$ for optical images and synthetic aperture radar (SAR) images using four classic CNN models, namely AlexNet [11], VGG-16 [19], ResNet-18 [7], and ResNet-34 [7]. In addition, we implemented the human interaction evaluation method described in Section 3.2. The results are presented in Fig. 4. Notably, the confidence differential curves of human evaluations resemble spikes in shape, suggesting that humans can quickly recognize specific objects after sufficient interactions. Additionally, introducing more interactions does not significantly improve classification performance. It is worth noting that humans require fewer interactions to reach maximum confidence for optical images than SAR images. Specifically, the peak of the confidence differential curve for optical images occurs at a lower order than that for SAR images, as shown in Fig. 3. This observation aligns with our intuition, as humans possess prior knowledge of various scenes in optical vision, making it easier to recognize objects in optical images than in SAR images.

In comparison, our results, presented in Figure 4, demonstrate that CNNs understand optical and SAR images in completely different ways. Specifically, we observe that for optical images, CNNs are better at encoding simple and complex interactions rather than intermediate interactions, as investigated in [3]. But for SAR images, CNNs are good at encoding intermediate interactions instead of too simple and complex interactions. This is an interesting phenomenon as CNNs become more similar to humans when used in SAR ATR. We hypothesize that this is because the targets in SAR images typically have a simple background, and therefore only when sufficient interactions are provided can CNNs recognize the targets. However, in the case of optical images, several patches of background can also provide sufficient discriminative information to classify a specific object. For example, the presence of waves in an image may indicate the presence of a whale instead of a lion.

Moreover, we found that the four CNN models show different trends in how interactions change with the order. For example, AlexNet can still capture some new information from complex interactions, while VGG-16, ResNet-18, and ResNet-34 almost cannot get any new information from complex interactions. We speculate that this could be because AlexNet has fewer layers and parameters to aggregate information from SAR images, and therefore needs to see more patches in the image to be fully confident in classifying its category.

5. Conclusion

In this study, we aim to provide an analytical interpretation of the cognitive gap between optical images and SAR images in CNNs using multi-order interactions. Additionally, we propose a subjective evaluation strategy to measure human interactions for different masked images. Experimental results demonstrate that CNNs' cognitive mechanism in SAR ATR is more similar to humans, as they exhibit better performance in encoding simple and complex interactions, contrary to intermediate interactions that are more important for optical images. Our research provides insights into the representation capacity of CNNs for SAR images, which can enhance the interpretability of CNNs in SAR ATR. This is an important step towards bridging the gap between CNNs and SAR ATR, which is crucial in geoscience and remote sensing applications.

Data Availability Statements

Dataset ILSVRC can be downloaded from the website https://www.image-net.org/challenges/LSVRC/.
MSTAR dataset can be downloaded from the website https://www.sdms.afrl.af.mil/index.php?collection=mstar

Acknowledgments

This work is funded by the National Natural Science Foundation of China (Grant No. 62276204, 61871301, 62071349).

References

[1] Binder, A., Montavon, G., Lapuschkin, S., Müller, K.R., Samek, W., 2016. Layer-wise relevance propagation for neural networks with local renormalization layers, in: Artificial Neural Networks and Machine Learning–ICANN 2016: 25th International Conference on Artificial Neural Networks, Barcelona, Spain, September 6-9, 2016, Proceedings, Part II 25, Springer. pp. 63–71.

[2] Cai, J., Zhang, L., Dong, J., Wang, C., Liao, M., 2022. Polarimetric sar pixel offset tracking for large-gradient landslide displacement mapping. International Journal of Applied Earth Observation and Geoinformation 112, 102867.

[3] Deng, H., Ren, Q., Zhang, H., Zhang, Q., 2022. Discovering and explaining the representation bottleneck of dnns, in: 2022 International Conference on Learning Representations (ICLR).

[4] Feng, Z., Cui, X., Ji, H., Zhu, M., Stanković, L., 2023. Vs-cam: Vertex semantic class activation mapping to interpret vision graph neural network. Neurocomputing .

[5] Fiori, S., 2003. Overview of independent component analysis technique with an application to synthetic aperture radar (sar) imagery processing. Neural Networks 16, 453–467.

[6] Fornaro, G., Verde, S., Reale, D., Pauciullo, A., 2014. Caesar: An approach based on covariance matrix decomposition to improve multibaseline–multitemporal interferometric sar processing. IEEE Transactions on Geoscience and Remote Sensing 53, 2050–2065.

[7] He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770–778.

[8] Huang, Y., Huang, H., 2023. Stacked attention hourglass network based robust facial landmark detection. Neural Networks 157, 323–335.

[9] Janizek, J.D., Sturmfels, P., Lee, S.I., 2021. Explaining explanations: Axiomatic feature interactions for deep networks. The Journal of Machine Learning Research 22, 4687–4740.

[10] Kaadoud, I.C., Bennetot, A., Mawhin, B., Charisi, V., Díaz-Rodríguez, N., 2022. Explaining aha! moments in artificial agents through ike-xai: Implicit knowledge extraction for explainable ai. Neural Networks 155, 95–118.

[11] Krizhevsky, A., Sutskever, I., Hinton, G.E., 2012. Imagenet classification with deep convolutional neural networks, in: Pereira, F., Burges, C., Bottou, L., Weinberger, K. (Eds.), Advances in Neural Information Processing Systems, Curran Associates, Inc.

[12] Lundberg, S.M., Erion, G.G., Lee, S.I., 2018. Consistent individualized feature attribution for tree ensembles. arXiv preprint arXiv:1802.03888 .

[13] Park, J.I., Park, S.H., Kim, K.T., 2012. New discrimination features for sar automatic target recognition. IEEE Geoscience and Remote Sensing Letters 10, 476–480.

[14] Peng, B., Peng, B., Zhou, J., Xie, J., Liu, L., 2022. Scattering model guided adversarial examples for sar target recognition: Attack and defense. IEEE Transactions on Geoscience and Remote Sensing 60, 1–17.

[15] Petsiuk, V., Das, A., Saenko, K., 2018. Rise: Randomized input sampling for explanation of black-box models. arXiv preprint arXiv:1806.07421 .

[16] Quan, S., Xiong, B., Zhang, S., Yu, M., Kuang, G., 2016. Adaptive and fast prescreening for sar atr via change detection technique. IEEE Geoscience and Remote Sensing Letters 13, 1691–1695.

[17] Ribeiro, M.T., Singh, S., Guestrin, C., 2016. " why should i trust you?" explaining the predictions of any classifier, in: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, pp. 1135–1144.

[18] Scalzo, B., Stanković, L., Daković, M., Constantinides, A.G., Mandic, D.P., 2023. A class of doubly stochastic shift operators for random graph signals and their boundedness. Neural Networks 158, 83–88.

[19] Simonyan, K., Zisserman, A., 2014. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 .

[20] Stanković, L., Mandic, D., 2023. Convolutional neural networks demystified: A matched filtering perspective-based tutorial. IEEE Transactions on Systems, Man, and Cybernetics: Systems .

[21] Vasile, G., Trouvé, E., Petillot, I., Bolon, P., Nicolas, J.M., Gay, M., Chanussot, J., Landes, T., Grussenmeyer, P., Buzuloiu, V., et al., 2008. High-resolution sar interferometry: Estimation of local frequencies in the context of alpine glaciers. IEEE Transactions on Geoscience and Remote Sensing 46, 1079–1090.

[22] Yang, R., Hu, Z., Liu, Y., Xu, Z., 2019. A novel polarimetric sar classification method integrating pixel-based and patch-based classification. IEEE Geoscience and Remote Sensing Letters 17, 431–435.

[23] Zhang, D., Zhang, H., Zhou, H., Bao, X., Huo, D., Chen, R., Cheng, X., Wu, M., Zhang, Q., 2021a. Building interpretable interaction trees for deep nlp models, in: Proceedings of the AAAI Conference on Artificial Intelligence, pp. 14328–14337.

[24] Zhang, H., Xie, Y., Zheng, L., Zhang, D., Zhang, Q., 2021b. Interpreting multivariate shapley interactions in dnns, in: Proceedings of the AAAI Conference on Artificial Intelligence, pp. 10877–10886.

[25] Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A., 2016. Learning deep features for discriminative localization, in: Proceedings of the 2016 IEEE conference on computer vision and pattern recognition (CVPR), pp. 2921–2929.