EXPLAINABILITY ANALYSIS OF CNN IN DETECTION OF VOLCANIC DEFORMATION SIGNAL

Teo Beker[1], Homa Ansari[1], Sina Montazeri[1], Qian Song[1], Xiao Xiang Zhu[1, 2]

[1] Remote Sensing Technology Institute (IMF), Weßling, German Aerospace Center (DLR), Germany
[2] Data Science in Earth Observation (SiPEO), Technical University of Munich (TUM), Munich, Germany

ABSTRACT

With improvement in the processing of synthetic aperture radar interferometry (InSAR) data, the detection of long-term volcanic deformations becomes possible. While deep learning (DL) models are considered black-box models, challenging to debug, the advances in explainable AI (XAI) help understand the model and how it makes decisions. In this paper, the model is trained on synthetic InSAR velocity maps to detect slow, sustained deformations. XAI tools, including Grad-CAM and t-SNE, are utilized for understanding and improving the trained model. Grad-CAM helps identify the slope-induced signal and salt lake patterns responsible for the model's misclassifications. T-SNE feature representation visualizations are used to estimate data sets and model class separation ability. Additionally, a sensitivity analysis shows the model performance with different intensity deformation data and uncovers the minimal detectable deformations of 1 cm cumulative deformation over five years.

Index Terms— Explainable AI, Grad-CAM, Volcano Detection, InSAR, Sensitivity Analysis

1. INTRODUCTION

Tracking cataclysmic events is of high importance to people. Synthetic Aperture Radar (SAR) satellites allow frequent measurements of deformations on a global level, thus, enabling deep learning (DL) to be applied to SAR Interferometry (InSAR) time-series for volcanic deformations detection [1]. In addition, advances in InSAR velocity map processing make it possible to distinguish mm/yr level deformations, facilitating more sensitive monitoring of volcanic deformations. However, to the author's knowledge, there has not been a previous attempt to use long-term velocity maps to detect slow, sustained deformations.

DL in volcanic deformation detection using InSAR has been applied to detect short and long-term deformations. In [2] it is demonstrated that short-term large-scale volcanic deformations can be detected using Alexnet architecture. Time-series simulated data were used to train an encoder-decoder architecture to filter out the noise [3]. In [4], the noise is filtered out of a time series by using Independent Component Analysis (ICA) algorithm, and in [5] the results are improved by developing the ICASAR algorithm.

As the models become common and the community is more acquainted with the concept, the (XAI) is given more importance. While DL models are considered black-box, explainable AI (XAI) practices help in debugging and make them more understandable. According to [6], out of the compared techniques, Occlusion, Grad-CAM and Lime are the most interpretable and reliable XAI methods. However, they have not been applied in volcanic deformation detection. Among the techniques, Grad-CAM is the least computationally expensive, and it shows which patterns the model focuses on and thus helps debug the model.

In this paper, we demonstrate how visual local and global "post-hoc" XAI tools make the trained model for volcanic deformations detection in the InSAR velocity map explainable. First, the Grad-CAM approach has been implemented to see which features are essential for model predictions. Then, t-SNE visualization of the extracted feature by the trained network has been used to compare test and fine-tuning data distributions. Finally, the sensitivity analysis is used to find the smallest detectable volcanic deformations. The explanations received in this way were helpful to improve the model, as it is presented hereinafter.

2. METHODOLOGY

Previously we trained a binary CNN classification network to determine whether the input image is a volcanic deformation map using simulated data. However, when tested on real data, the accuracy drops significantly. Thus, three XAI tools are introduced in this section to explain the trained model and identify the problem. For an explainability analysis, different aspects are considered:

- Gradient Class Activation Mapping (Grad-CAM) is used to check if the model focuses on correct features in the input images when making decisions.
- The feature representation analysis using t-distributed stochastic neighbor embedding (t-SNE) is performed to compare the fine-tuning set with the real test set.
- The sensitivity analysis is performed by scaling the real test set images by selected temporal baselines and wrapping them to see how small deformations the model can detect.

2.1. Grad-CAM

Grad-CAM [7] is a local post-hoc feature attribution saliency map generation technique. It tracks the propagated gradient to show where the model is "looking" when deciding. Unlike its predecessors, Grad-CAM is general and allows the use on any model or any layer of the model.

As the XAI methods get more researched and compared, Grad-CAM is getting more recognition. While many feature attribution methods suffer from the independence of model randomization and label permutation, Grad-CAM does not [8]. According to [6], among tested XAI approaches, it is one of the three most reliable methods to be used in Earth observation.

This paper uses XAI to understand the model's decisions and identify why the model performance on the real test set was suboptimal.

2.2. T-SNE Feature Representations

T-SNE is a widely-used unsupervised dimensionality reduction algorithm [9]. In this paper, the CNN-extracted hyper-dimensional features of fine-tuning and test sets are input to t-SNE. Then, the received embedded two-dimensional (2D) features are used for visualization.

In CNNs, usually, convolutional layers are regarded as feature extractors. Outputs of the last convolutional layer (the second last dense layer in this paper) can be seen as the extracted hyper-dimensional features of the input images. It is vital to make sure that the extracted features of the training set are consistent with that of the test set for the same category. However, the high dimensionality hinders the understanding of the features. Thus, the extracted features of different sets are embedded for visualization purposes using t-SNE.

2.3. Sensitivity Analysis

Sensitivity analysis shows us the smallest detectable volcanic deformation and how the model performance changes by changing the data's intensity. The sensitivity analysis tests the model performance on the scaled test sets according to the temporal baselines from 1 - 6 years. Since the model is trained on cumulative deformations over five years, this can also be understood as a change of wrapping wavelength (1.11cm - 6.65cm). Because of wrapping, this approach exposes the model to differently looking data, effectively creating more diverse real test samples. As there are only ten deforming volcanoes in our real data test set, sensitivity analysis tests the model on a broader range of scales of deformations. It allows us to determine the model's performance more broadly and robustly. It also enables us to find the smallest detectable deformation by our model. This approach is not taking the signal-to-noise ratio (SNR) of the deformation and residual atmospheric noise and slope-induced signal into the analysis, which can be considered the same for an image changing the temporal baseline.

To evaluate the model sensitivity, we compare the metrics of recall, precision, and FPR. Besides the metrics, the number of detected frames with volcanic deformations will be recorded, and the scale of the minimal detected deformations.

3. RESULTS

3.1. Data and Model

The velocity maps [10] extracted from large interferometric stacks spanning about five years of the region of the central South American Andes were used. The complete real data are reserved as the test set. A real test set was created by cropping velocity maps around volcanoes and transforming them into cumulative deformations over five years. The cumulative deformations are wrapped between $-\pi$ and π [1], given the Sentinel-1 wavelength of ≈ 5.547 cm. The spatial resolution of 200m and extent of 102.4 km by 102.4 km was used to create the frames. The frames were patched using 75% overlap. The test set is highly imbalanced given a limited number of deforming volcanoes. Since the real data after the creation of the velocity maps are scarce, the simulations of the residual noise [4], and volcanic deformations [11] were used to create a balanced synthetic training (297,752 samples) and validation set (33,082 samples).

After comparing six different model architectures on the real test set, InceptionResNet v2 was chosen as the classifier for volcanic deformation detection, as it gave the best performance by a margin. Nevertheless, none of the models gave satisfactory results, the best model achieving 58% area under the curve receiver operating characteristic (AUC ROC).

3.2. Unaccounted Patterns

The saliency maps are plotted for three real velocity maps in Fig 1 using the Grad-CAM technique. In some samples, Grad-CAM scores in areas with salt lakes and slope-induced signals are significantly high. Furthermore, the model predicted the first two samples as volcanic deformations with high confidence (65.35% and 100%), implying that the model confused salt lakes and slope-induced signals for volcanic deformations. In the bottom case, it is correctly classified for the wrong reasons. The right column shows more feasible Grad-CAM maps and classification probabilities achieved after fine-tuning.

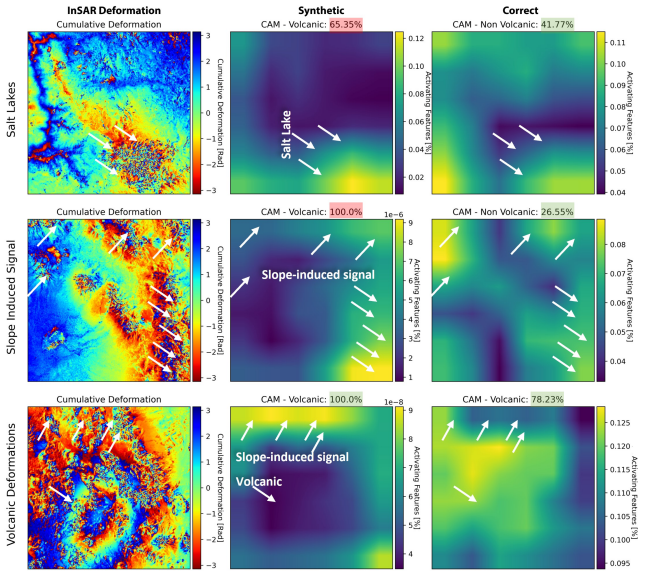

Fig. 1: The velocity maps are on the left, and the Grad-CAM maps of the model trained on synthetic data and fine-tuning set, are shown in the middle and on the right.

The results of Grad-CAM suggest that it is necessary to fine-tune the model with additional data. Therefore, we trained the last layer of the model with a hybrid synthetic-real set.

The fine-tuning set was extracted from the mountainous region around the volcanoes. It did not include any area used by the real test set but always collected samples close around it. This way, the extracted slope-induced signal resembled the one in the real test set. 836 patches were extracted this way. Half was used as a non-volcanic deformation class, and to the other half, simulated volcanic deformations were added to form a volcanic deformation class.

The fine-tuned models performed significantly better than the original synthetic model, as is represented by the increase from 58% to 86% AUC ROC. Besides, based on the saliency maps (right column in Fig. 1) of the same three velocity maps created by Grad-CAM on the fine-tuned model, the model now paid more attention to the volcanic deformation patterns.

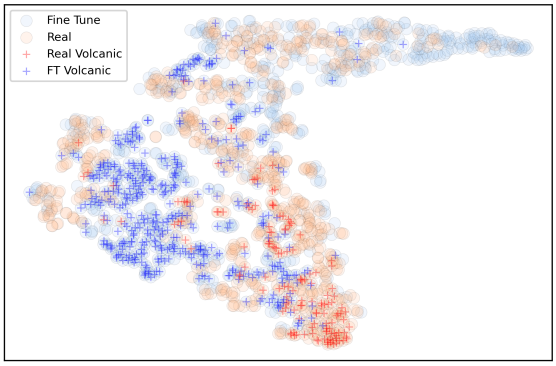

Fig. 2: The comparison of samples of real and fine-tuning set visualized using t-SNE. The fine-tuning data are different to real data, but there is a significant overlap of distributions.

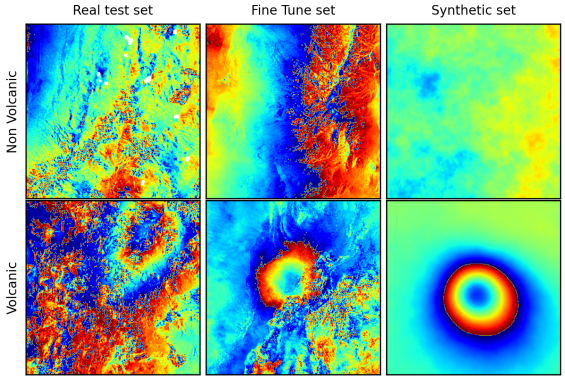

Fig. 3: Comparison of the data sets used. The synthetic set did not account for slope-induced patterns. Therefore, it is similar only to low-pass filtered real data. Meanwhile, the fine-tuning set is similar to the real test set. Note: The test set is the only set using real velocity maps.

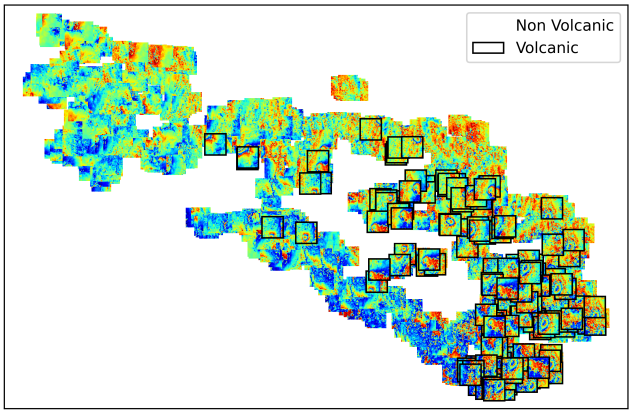

Fig. 4: T-SNE transformation of feature space of FT4 model on the real test data, showing the sample with and without volcanic deformations. It is noticeable that the majority of the volcanic deformations are grouped well together, while there are smaller clusters and about 10 examples of partial deformations which are further away from these clusters.

3.3. Similarity of Data Sets and Separability of Classes

First, the fine-tuning and real test set feature representations were compared. The features were extracted from the fine-tuned model's second last, flattening layer. The layer contains 1536 features. These features represent the high-level visual features of the images. The features were coded to two dimensions using the t-SNE model, which can be seen in Fig 2. Visual analysis shows that while the fine-tuning set is similar to the real data set, it does have a slightly different distribution. Most of the fine-tuning set volcanic deformations are grouped, containing only a couple of the real test set volcanic deformations. The same goes for the real test set volcanic deformations. That means that most volcanic samples are noticeably different between the sets. Both sets cover a large area of slope-induced signal, but a couple of clusters are not accounted for by the fine-tuning set, pointing to the imperfections of the fine-tuning set. The improvements over the initial training set can be seen in Fig 3.

We can check how the images are grouped or separated by the fine-tuned model in Fig 4. This is a visual confirmation that the images are grouped well by the visible patterns. It is also noticeable that besides the patterns, the intensity of the deformation plays a significant role in grouping the images.

3.4. Sensitivity Analysis

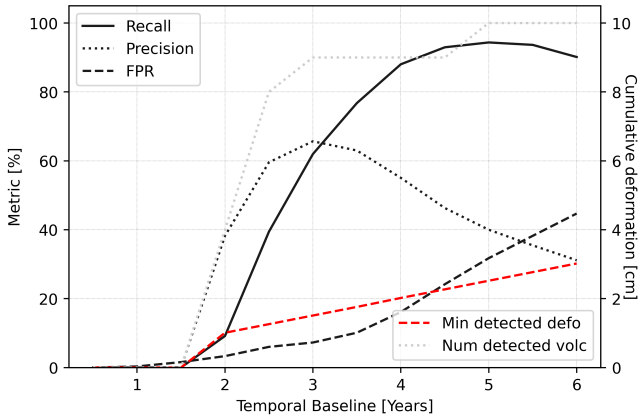

Fig. 5: Sensitivity analysis - testing the model performance using the different temporal baselines of the real test set data. Precision increases up to three years and declines afterwards. This happens for the class imbalance, only about 20% of the real test set is positive, and therefore as soon as the FPR starts increasing at faster rate the Precision drops. Recall peaks at 5 years, which is the time period the model was trained for. The minimal volcanic deformation detection threshold moves with the increase of the temporal baseline, starting with 1 cm at temporal baseline of 2 years. The number of detected volcanoes shows if every used frame containing volcanic deformations has at least one patch flagging it as positive.

Finally, to estimate the performance of the model detection capabilities, we checked the scale of the volcanic deformations which the model can detect. Since the real test set was limited, we performed a sensitivity analysis by scaling the temporal baseline of the data from 1-6 years. The model was trained on the ~ 5 year temporal baseline. Exposing the model to the different temporal baselines shows how the model performs with the increased or decreased intensity of the signal while exposing it to data looking slightly different from the

original (because of the wrapping process). The results can be seen in Fig 5.

The results show that the minimal detected volcanic deformation is about 1 cm. The same detected volcano, Cordon Del Azufre, stays the minimal detected deformation throughout the temporal baselines. First, correct classifications happen from two year temporal baseline, thus implying there is a threshold for the smallest detection, even though the SNR stays about the same.

Precision is the highest at the temporal baseline of 3 years, after which it starts falling off with the increase of the FPR. Also, nine out of ten frames containing volcanic deformations are detected at the same baseline. The late detection of the tenth frame can be explained by it containing only the edge of deformation (not containing the center of the volcano) of the Sabancaya volcano. The recall is highest at five years, the model's original baseline, and it declines afterward as the noise signal becomes pronounced and wrapping creates more edges in the images.

4. CONCLUSION

In this paper, the XAI tools are demonstrated to help improve the DL model and understand its decisions and why it makes them. The Grad-CAM analysis identifies the presence and significance of the slope-induced signal and salt lakes in the real test set. With this knowledge, the model is fine-tuned to improve its performance.

The feature representations analysis compares the fine-tuning set to the real set. While the fine-tuning set is quite similar to the real test set, there are detected differences in distributions. Feature representations are also used to examine the ability of the model to separate the volcanic deformations from the other patterns in the data.

Finally, the sensitivity analysis is performed to determine the model's performance with a change of the temporal baseline. The minimal detectable deformation is identified, measuring 1cm at two year baseline. All of the volcanoes whose center is contained in the data can already be detected at three years temporal baseline, which is also the point giving the highest precision and relatively low FPR.

Using explainability to guide the modeling process can ease the work with the black-box models and give significant insights into the models' decision-making process. This work brings insights to improve the model performance and suggest future steps to improve the fine-tuning sets.

5. ACKNOWLEDGEMENTS

The paper is based on research performed as a part of TecVolSA project [12]. The data has been prepared by InSAR team on DLR and special thanks to Robert Shau for patience and help with data. Also great thanks go to GFZ and especially to Rene Mania for simulation of volcanic deformations and Tomas Walter for their help with the domain knowledge on volcano deformations and regional geologic processes. For support with their discussions and advice on the DL we thank Yuanyuan Wang, Syed Mohsin Ali, and Ivica Obadić.

6. REFERENCES

[1] Nantheera Anantrasirichai, Juliet Biggs, Fabien Albino, and David Bull, "The application of convolutional neural networks to detect slow, sustained deformation in insar time series," *Geophysical Research Letters*, vol. 46, no. 21, pp. 11850–11858, 2019.

[2] Nantheera Anantrasirichai, Fabien Albino, P Hill, D Bull, and Juliet Biggs, "Detecting volcano deformation in insar using deep learning," *arXiv preprint arXiv:1803.00380*, 2018.

[3] Jian Sun, Christelle Wauthier, Kirsten Stephens, Melissa Gervais, Guido Cervone, Peter La Femina, and Machel Higgins, "Automatic detection of volcanic surface deformation using deep learning," *Journal of Geophysical Research: Solid Earth*, vol. 125, no. 9, pp. e2020JB019840, 2020.

[4] ME Gaddes, A Hooper, M Bagnardi, H Inman, and F Albino, "Blind signal separation methods for insar: The potential to automatically detect and monitor signals of volcanic deformation," *Journal of Geophysical Research: Solid Earth*, vol. 123, no. 11, pp. 10–226, 2018.

[5] ME Gaddes, Andy Hooper, and Marco Bagnardi, "Using machine learning to automatically detect volcanic unrest in a time series of interferograms," *Journal of Geophysical Research: Solid Earth*, vol. 124, no. 11, pp. 12304–12322, 2019.

[6] Ioannis Kakogeorgiou and Konstantinos Karantzalos, "Evaluating explainable artificial intelligence methods for multi-label deep learning classification tasks in remote sensing," *arXiv preprint arXiv:2104.01375*, 2021.

[7] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 618–626.

[8] Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim, "Sanity checks for saliency maps," *arXiv preprint arXiv:1810.03292*, 2018.

[9] Laurens Van der Maaten and Geoffrey Hinton, "Visualizing data using t-sne.," *Journal of machine learning research*, vol. 9, no. 11, 2008.

[10] Homa Ansari, Francesco De Zan, Giorgio Gomba, and Richard Bamler, "Emi: Efficient temporal phase estimation and its impact on high-precision insar time series analysis," in *IGARSS 2019-2019 IEEE International Geoscience and Remote Sensing Symposium*. IEEE, 2019, pp. 270–273.

[11] Mehdi Nikkhoo, Thomas R Walter, Paul R Lundgren, and Pau Prats-Iraola, "Compound dislocation models (cdms) for volcano deformation analyses," *Geophysical Journal International*, p. ggw427, 2016.

[12] Sina Montazeri, Homa Ansari, Francesco De Zan, René Mania, Robert Shau, Teo Beker, Alessandro Parizzi, Mahmud Haghshenas Haghighi, Peter Niemz, Simone Cesca, et al., "Tecvolsa: Insar and machine learning for surface displacement monitoring in south america," in *EGU General Assembly Conference Abstracts*, 2021, pp. EGU21–6086.