Explainable Artificial Intelligence for Face Presentation Attack Detection

Wilson Silva[1,2]
wilson.j.silva@inesctec.pt
João Ribeiro Pinto[1,2]
joao.t.pinto@inesctec.pt
Tiago Gonçalves[1,2]
tiago.f.goncalves@inesctec.pt
Ana F. Sequeira[1]
ana.f.sequeira@inesctec.pt
Jaime S. Cardoso[1,2]
jaime.cardoso@inesctec.pt

[1] INESC TEC
Porto, Portugal
[2] Faculty of Engineering, University of Porto
Porto, Portugal

Abstract

The use of deep learning techniques for face presentation attack detection (PAD) is increasingly common due to their ability to reach strong accuracy performances. Nonetheless, the use of complex models such as the ones produced with deep learning techniques raises safety and trust concerns, as one is not able to understand the motifs behind model decisions. Furthermore, traditional metrics of evaluation fall short in terms of capturing the desirable working properties of models, which is particularly worrisome when working in high-regulated areas, such as biometrics. In this work, we propose the use of interpretability techniques to further assess the robustness of face PAD models. Moreover, we also define desirable properties for a face PAD model to have, which can be evaluated through interpretability. Experiments were performed using the ROSE Youtu video collection and showed the additional value of interpretability in the identification of model robustness.

1 Introduction

Nowadays, deep learning algorithms are excelling in most of the artificial intelligence (AI) fields, including in the biometrics and forensics domain. Although these models can indeed achieve incredible performances due to their complexity and flexibility, it is also true that sometimes these performances are obtained by a focus in wrong/biased information instead of domain significant information [4]. Therefore, an evaluation performed based on only traditional metrics may be misleading, making us trust a model that is not robust enough to be deployed in the real-world.

With regards to face PAD, the use of deep learning techniques is also increasingly common [6]. Furthermore, the diversity of presentation attacks that can happen in a real-world scenario increase the importance of checking the robustness of the deep models, as they may focus on attack-specific or spurious information instead of more general features capable of characterising what an attack means [3].

To overcome the limitations of evaluating a face PAD model only with the traditional metrics, we propose in this work the use of interpretability methods to further assess how robust is a model, by checking which information is determining the deep learning model decision. Interpretability or explainability (we use both terms interchangeably) is the process of understanding which features, or which process, led to the machine learning model decision. Doshi-Velez and Kim categorised these techniques into three different groups, namely, pre-, in-, and post-model [2]. In the last years, interpretability research has focused attention on the in- and post-model interpretability groups, i.e., in the proposal of interpretable models by design [9], or in the proposal of interpretability methods to analyse previously built models [1].

In this work, we also assess the fulfilment of important properties defined by Sequeira *et al.* [8], such as (1) explanations for the same sample should be similar whether or not it is seen during training (data swap); and (2) explanations for the same sample should be similar whether or not the model is trained to detect that specific attack (One-Attack vs. Unseen-Attack).

2 Methodology

A presentation attack detection method receives as input a biometric trait measurement and returns as output a prediction of the classification of that measurement as belonging to a living individual (*bona fide*) or as being a spoof attempt to intrude the system (*attack*). In this work, our method consists of an end-to-end convolutional neural network, with its architecture being described in Figure 1. Since the focus of the work is the study on the interpretability of the face PAD model, we chose a relatively simple architecture.

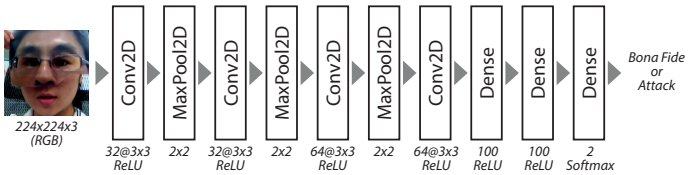

Figure 1: Architecture of the implemented PAD model.

With regards to the interpretability method to be used in this work, we selected the well-known Grad-CAM method [7], as it has the flexibility to generate explanations for any layer of the network, and also allows us to obtain class-specific explanations.

3 Experimental Assessment

The experiments were performed with the ROSE-Youtu Face Liveness Detection Dataset [5], which is composed of 3497 videos acquired from twenty different subjects. For each subject there are several "genuine", and "attack" videos (types of attack, and number of frames extracted are presented in Table 1). The PAD model previously presented was implemented in Keras and trained for 150 epochs with early-stopping (based on validation loss). To avoid overfitting, regularization techniques such as dropout and data augmentation were used.

Table 1: Characteristics of the presentation attack instruments in the ROSE Youtu dataset (N.I. stands for "number of images", *i.e.*, frames extracted from the videos).

Attack	Type of presentation attack instruments	N.I.
-	Genuine (*bona fide*)	2794
#1	Still printed paper	1136
#2	Quivering printed paper	1188
#3	Video of a Lenovo LCD display	923
#4	Video of a Mac LCD display	1113
#5	Paper mask without cropping	1194
#6	Paper mask with two eyes and mouth cropped out	608
#7	Paper mask with the upper part cut in the middle	1162

The quantitative results in terms of *Bona fide Presentation Classification Error Rate (BPCER)*, *Attack Presentation Classification Error Rate (APCER)*, and *Equal Error Rate (EER)* are shown in Table 2. As illustrated in Table 2, we performed the experiments using two different evaluation frameworks: *one-attack* (model is trained and tested with only one type of attack), and *unseen-attack* (model is trained with all but one attack, and tested with the remaining attack). Even though the focus of the work is not on the performance of the face PAD model, the method's performance is in line with state-of-the-art methods. The results with regards to the Unseen-Attack framework are worse than the ones related to the One-Attack framework, which indicate model generalization problems.

Table 2: PAD performance of the models for One-Attack and Unseen-Attack evaluation frameworks. (EER, APCER, and BPCER in %)

Attack	One-Attack			Unseen-Attack		
	EER	*APCER*	*BPCER*	*EER*	*APCER*	*BPCER*
#1	7.29	12.15	3.06	5.90	6.94	4.90
#2	3.62	6.67	1.35	5.55	**3.00**	10.65
#3	2.79	8.37	0.12	10.38	26.29	4.28
#4	12.66	30.38	1.84	25.34	45.73	**3.92**
#5	1.61	**1.61**	1.59	**4.84**	3.55	7.10
#6	4.46	5.10	1.10	10.19	12.74	7.71
#7	**0.73**	5.23	**0.00**	15.49	34.31	7.71

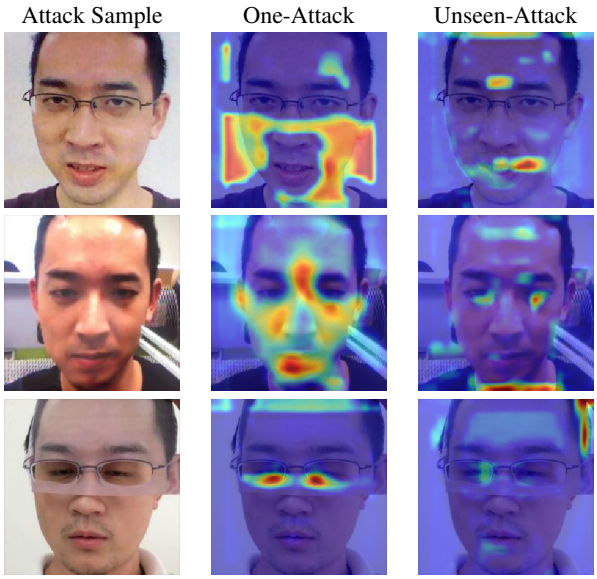

Figure 2: Explanations for correctly classified attack samples (TP) in the One-Attack (2nd column) or Unseen-Attack (3rd column) frameworks. Each row corresponds to one specific type of attack, top to bottom: #1, #4, and #7.

Apart from the usual quantitative evaluation performed for PAD models, we introduce here a qualitative evaluation of model properties based on explanations. With this regard, two types of experiments were performed: comparing explanations for the same attack sample when in the one-attack framework or the unseen-attack framework; and, comparing explanations when attack samples of a random subject are present in train or test (swap experiment). The results obtained with these two approaches are presented in Fig. 2 and Fig. 3. As it can be observed in Fig. 2, the explanations generated for the same samples in the *one-attack* and *unseen-attack* frameworks are quite different, not showing coherence on the information that is relevant to making the decision, which again indicates there are generalization issues with the models. On the other hand, the models demonstrated to be robust with regards to unseen subjects, as the explanations generated in the swap experiment show relevance of the same regions independently of the subject under analysis being in train or test.

4 Conclusions and Future Work

In this work, interpretability techniques were explored to further assess the robustness of face PAD models. Moreover, we studied several desirable properties for a face PAD model to fulfil that are only verifiable through an interpretability analysis of the models. Nonetheless, this interpretability evaluation can only be done qualitatively, therefore, lacking objectivity. In future work, we aim to find ways of quantifying the information obtained with the interpretability analysis.

Acknowledgements

This work was financed by the ERDF – European Regional Development Fund through the Operational Programme for Competitiveness and Internationalization - COMPETE 2020 Programme and by National Funds through the Portuguese funding agency, FCT – Fundação para a Ciência e a Tecnologia within project "POCI-01-0145-FEDER-030707", and within the PhD grants "SFRH/BD/137720/2018", "SFRH/BD/139468/2018" and "SFRH/BD/06434/2020".

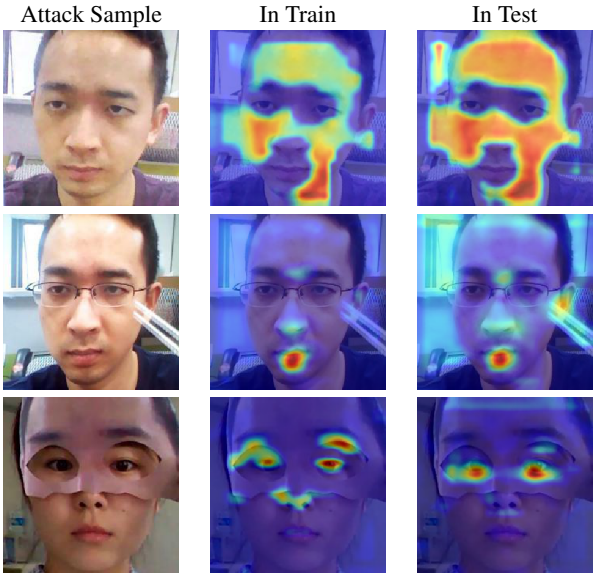

Figure 3: Grad-CAM Explanations for correctly classified attack samples when a subject is in the train set (2nd column) or in the test set (3rd column). Each row corresponds to one specific type of attack, top to bottom: #1, #4, and #7.

References

[1] Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one*, 10(7):e0130140, 2015.

[2] Finale Doshi-Velez and Been Kim. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*, 2017.

[3] Pedro M Ferreira, Ana F Sequeira, Diogo Pernes, Ana Rebelo, and Jaime S Cardoso. Adversarial learning for a robust iris presentation attack detection method against unseen attack presentations. In *2019 International Conference of the Biometrics Special Interest Group (BIOSIG)*, pages 1–7. IEEE, 2019.

[4] Sebastian Lapuschkin, Alexander Binder, Grégoire Montavon, Klaus-Robert Muller, and Wojciech Samek. Analyzing classifiers: Fisher vectors and deep neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2912–2920, 2016.

[5] Haoliang Li, Wen Li, Hong Cao, Shiqi Wang, Feiyue Huang, and Alex C Kot. Unsupervised domain adaptation for face anti-spoofing. *IEEE Transactions on Information Forensics and Security*, 13(7): 1794–1809, 2018.

[6] Daniel Pérez-Cabo, David Jiménez-Cabello, Artur Costa-Pazo, and Roberto J López-Sastre. Deep anomaly detection for generalized face anti-spoofing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019.

[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*, pages 618–626, 2017.

[8] Ana F Sequeira, Wilson Silva, João Ribeiro Pinto, Tiago Gonçalves, and Jaime S Cardoso. Interpretable biometrics: Should we rethink how presentation attack detection is evaluated? In *2020 8th International Workshop on Biometrics and Forensics (IWBF)*, pages 1–6. IEEE, 2020.

[9] Wilson Silva, Kelwin Fernandes, and Jaime S Cardoso. How to produce complementary explanations using an ensemble model. In *2019 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2019.