Explainable Visualization for Morphing Attack Detection

Blind revision

Abstract: Detecting morphed face images has become critical for maintaining trust in automated facial image verification systems. Researchers have discovered that deep facial recognition systems intended for generalizability increased their vulnerability to exploitation. Attacks based on altered face images offer a significant security concern. Morphing can be understood as a technique to combine two or more look-alike facial images from one subject and an accomplice, who could apply for a valid passport by exploiting the accomplice's identity. Deep Neural Networks have demonstrated good performance in detecting morphed images. However, they lack transparency, and it is unclear how they differentiate between authentic and morphed facial photos. As a result, this phenomenon needs careful consideration for safety and security-related applications. This paper will explore Layer-wise Relevance Propagation (LRP) to determine the most relevant features. We fine-tune a VGG pre-trained network for face morphing attack detection. LRP is then used to investigate the decision-making processes and see what input pixels take part in the attack detection. This paper shows that CNN only considers a small part of the image, usually around the eyes, nose, and mouth.

Keywords: Face morph; Morph Attack Detection; Layer-wise Relevance Propagation; VGG19; Deep Neural Network; Convolutional Neural Network; APCER; BPCER; Confusion Matrix

1 Introduction

Face recognition systems (FRS) is a technology that enables an individual to be recognized based on their unique biological traits captured from a facial image [Ve21]. Given the strong generalization capabilities of such systems, an adversary can execute targeted attacks against FRS that use morphed face images, as presented by Ferrara et al. [FFM14]. Face morphing is the smooth transformation of two facial pictures into one. Morphing attacks have developed as a severe threat to enrollment in recent years, undermining facial recognition systems' capabilities. As a result, this attack violates the rule of exclusive ownership [Ve21]. Morphing attack detection (MAD) algorithms have succeeded in determining alterations using deep learning in recent years.

With the advancement of deep learning algorithms, biometric-based identification and verification have become a commonly utilized methodology for a variety of secure access control applications [Ve21]. Classification of images has become a critical component of a wide variety of computer vision applications, with nonlinear methods such as convolutional neural networks (CNNs) serving as the gold standard [La16]. While approaches based on learned features can reach a high level of accuracy, they act like black boxes [SHE21a]. As

neural networks become more widely used, the topic of how these models' conclusions may be interpreted becomes increasingly important [Ra21]. While precision is necessary for network performance, generality and robustness are equally critical. One aspect of neural networks is that they frequently employ only the data necessary to perform their task and reject additional helpful information. Worst case, a neural network learns to make correct decisions for the wrong reasons [Se20].

The discipline of explainable artificial intelligence has seen the development of a plethora of methodologies. Bach et al. [Ba15] proposed the concept of layer-wise relevance propagation, which has established itself as a notable method for enhancing the interpretability of CNNs. This explanatory approach generally examines the model's interpretability from a black-box perspective and will be utilized in this paper concerning MAD networks.

2 Background

Morphing can be understood as a technique to combine two or more look-alike facial images from one subject and an accomplice, who could apply for a valid passport by exploiting the accomplice's identity [Ve21]. This technique can be used to construct artificial biometric samples that mimic the biometric information, as seen in Figure 1. Such face morphing attacks have implications on the integrity of automatic and manual identity verification procedures, like those conducted at country borders [SHE21a].

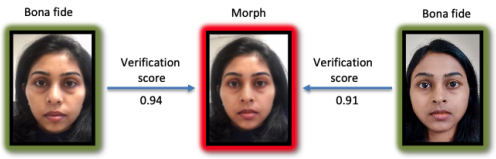

Fig. 1: Face morph illustrated in the middle getting a high similarity score against two bona fide samples from different individuals. Illustration is gathered from this paper [Ve21]

Since the morphing process alters the pixel positions, some mismatched pixels may result in noise-generating artefacts and ghost-like pictures, giving the photos an unreal aspect [Ve21]. After creating the morphed face image, it can be further processed and manipulated to remove or minimize these unnatural aspects. Automatically created morphs may introduce artefacts, which can be avoided if the attacker creates a single high-quality morph and manually optimizes the final image [Sc19]. In general, it is anticipated that mechanically created databases of morphed face photos will have a lower quality than real-world attack scenarios [Sc19].

2.1 Morphing Attack Detection

The MAD algorithms proposed thus far have been trained and evaluated on datasets with constrained distributions of image features and technological variety [Sc20]. The recent

NIST FRVT MORPH results [Ng20] indicate that most MAD algorithms submitted lack resilience and performance when applied to unknown and demanding datasets [Sc20].

Single Image Based MAD (S-MAD) is to detect a face morphing attack effectively using a single image supplied to the algorithm. The morphed image might be digital or re-digitized [Ve21]. S-MAD is a difficult task since it is supposed to be robust against differences in picture quality, multiple types of sensors, morph creation tools, and various print-scan procedures [Ve21, TB21].

Researchers have successfully used deep learning S-MAD algorithms to categorize bona fide and morphed images. Deep CNNs are utilized for the most part. Most previously published work uses pre-trained networks and transfer learning. Although deep CNNs outperform hand-crafted texture descriptor-based MAD algorithms on both digital and print-scan data, their generalizability and robustness are restricted [Ve21].

2.2 Explainability of deep learning models

Since deep learning is doing an excellent job in detecting morphing attacks, it is helpful to be able to explain what the algorithm uses in its decision-making. Visualization techniques are a common approach. Most approaches for face morph recognition are trained and evaluated on a single database utilizing a single morph generation algorithm [Sc19, TB21]. As a result, the training data must have a high degree of variance to avoid overfitting on database-specific artefacts [Sc20].

2.2.1 Layer-wise Relevance Propagation

Introduced by Bach et al. [Ba15], the Layer-wise Relevance Propagation (LRP) interpretability approach assigns significance to each pixel in the input image [SHE21a]. The LRP's mathematical foundation is built on a deep Taylor decomposition. It assigns relevance layer by layer, starting with a single selected neuron representing a single class and ending with the image via the CNN. Each layer's relevance is communicated backwards into the preceding one by a set of rules. We follow the rules currently regarded as best practice for LRP [SHE21b]. These are epsilon-decomposition for the fully connected layers. Alpha-beta - decomposition with $a = 2$, $b = -1$ and flat decomposition for the convolutions layer [SHE21b]. These criteria are intended to direct attention to the neurons in the prior layer that are required for each neuron in the current layer to fire [SHE21b].

LRP considers the CNN's overall structure, the classification component, and the convolutional layer activations and weights. The relevance is assigned so that regions that considerably contribute to the activation are given a positive value, illustrated with red colour. In contrast, areas that significantly inhibit its activation are assigned a negative value, presented with blue colour. This enables the production of finer heatmaps and assigning a

relevance score to each pixel, defining its ability to either contribute to or prevent activation [SHE21a].

3 Methodology

To study the explainability of the deep-learning based S-MAD algorithms, we first train a VGG19 network to classify morphed and bona fide images and then use LRP to interpret what has been learned by the model to make the classification. Due to the privacy reason, the size of the morphing dataset is usually limited. Hence, the VGG19 network in this work is fine-tuned based on weights pre-trained on ImageNet-1k dataset. After the fine-tuning, the model will be evaluated using standardized metrics ISO/IEC 30107-3 [IS17].

For the explanation of the model, we employ LRP [1] to gain insight into the decision-making process to interpret the MAD algorithm's accuracy and robustness [Se20]. More specifically, we use LRP to determine the input relevance in the bona fide and morphed images. Different kinds of explanatory photos are computed and visualized for discussions.

The dataset used in this paper is a subset of the database presented by Phillips et al. [Zh21]. The data originates from the FRGC-v2 dataset and consists of 140 unique participants from the FRGC-v2 collection based on the high-quality facial photos, where images similar to passport pictures were chosen. 47 of the 140 data individuals are female, whereas 93 are male. Each subject has a sample size of 7 to 21 images. The images are cropped by utilizing MTCNN presented by Zhang et al. [Zh16]. The morphed images were generated from the FRGC-v2 subset by using four different morphing algorithms.

The first morphing algorithm LMA [Ra17] 'Landmark-I' is mentioned in this paper [Zh21]. The 'Landmark-I' algorithm employs three different face morph generation techniques based on facial landmarks constrained by Delaunay triangulation with blending. The second morphing algorithm, LMA_UBO [FFM19] 'Landmark-II' [Zh21], is an improved landmark-based morphing algorithm. Including crop-pasting of the face region and other post-processing techniques to reduce the ghost artifacts from morphing. The third morphing algorithm is StyleGAN [Ve20]. Facial images are projected to the latent space of a pre-trained StyleGAN model, linearly averaged, and then synthesized by the generator. The last morphing algorithm used was the MIPGAN-I [Zh21], which improved the GAN-based morphing algorithm. Different loss functions, including identity-based loss function, are used to improve the morphing attack potentials.

4 Experiments & Results

Since MAD performance can be considered as a binary classification problem, the following metrics are widely used to benchmark the MAD algorithms. The performance of the detection algorithms is reported according to metrics defined in ISO/IEC 30107-3 [IS17].

[1] `https://github.com/fhvilshoj/TorchLRP`

The Attack Presentation Classification Error Rate (APCER) is defined as the proportion of attack samples incorrectly classified as bona fide images [Ve21]. The Bona fide Presentation Classification Error Rate (BPCER) is defined as the proportion of bona fide images incorrectly classified as a morphed image in the system [TB21].

The fine-tuned neural network that yielded the best results got a training accuracy of 0.994574 and a validation accuracy of 0.891724. Based on the results of this fine-tuned MAD algorithm, we used LRP to visualize what the neural network had used in its decision-making process. The bona fide images are presented in Figure 2 Top, and the morphed images are presented in Figure 2 Bottom. Table 1 contains the models performance metrics.

Tab. 1: Overview of APCER and BPCER and the values used to calculate them for the validation phase. Sorted by the 10 epochs with the highest accuracy score.

Epoch	**TP**	**TN**	**FP**	**FN**	**APCER**	**BPCER**	**Accuracy**
1	5288	4183	116	1034	0.026983	0.163556	0.891724
2	5260	4109	144	1108	0.033859	0.173995	0.882120
3	4400	4864	1004	353	0.171098	0.074269	0.872234
4	5356	3903	48	1314	0.012149	0.197001	0.871763
5	5192	4024	212	1193	0.050047	0.186844	0.867715
6	4900	4198	504	1019	0.107188	0.172157	0.856605
7	3988	4954	1416	263	0.222292	0.061868	0.841917
8	4368	4541	1036	676	0.185763	0.134021	0.838810
9	5400	3497	4	1720	0.001143	0.241573	0.837680
10	5376	3509	28	1708	0.007916	0.241107	0.836550

Figures 2 shows a small sub-sample of twelve bona fide and morphed pictures from the overall dataset. Using LRP, we see what went into the decision of the MAD algorithm with the highest accuracy when classifying the images as bona fide or morphs.

The visual results show that the neural network primarily focuses on the eyes, nose, and mouth. This is best illustrated in the patternnet explanations. The algorithm also takes into account some of the hair features, as well as the edges of the faces. The patterned explanation shows a pattern where most of the image negatively influences its decision, while some areas in the forehead and cheek positively influence its decision.

5 Discussion

It should be mentioned that reliable detection of face morphing attacks continues to be a challenge, and numerous open issues exist in the research field of MAD algorithms [Sc19]. One of these issues is the absence of large-scale publicly available datasets with more individual variation and a technological variation to reflect the real world [Ve21]. In addition, generating high-quality face morphs automatically continues to be complicated. The dataset used in this paper arguably has more artefacts and ghosting in the morphed images than an attacker would achieve manually for a specific purpose.

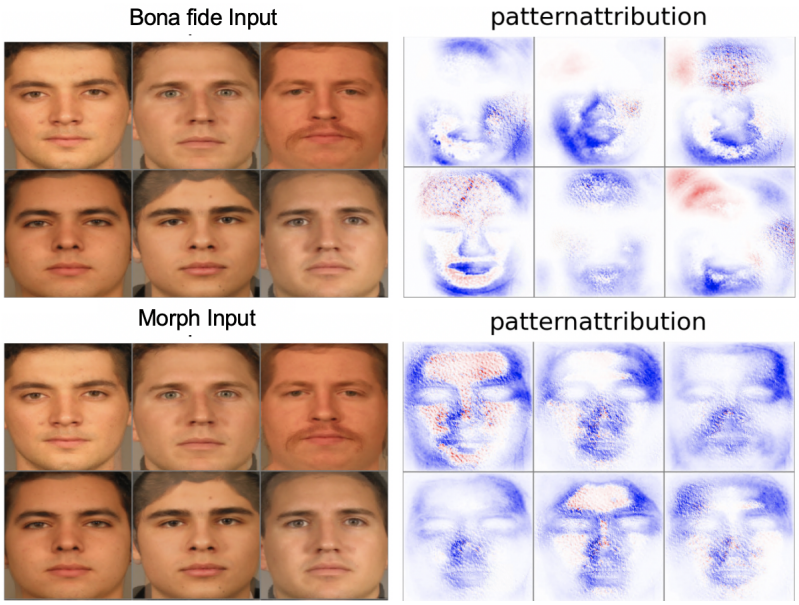

Fig. 2: Visualisation using the patternnet explanation. Top: bona fide images. Bottom: Morph images.

From Table 1, the validation accuracy during the top ten epochs during fine-tuning is relatively high. However, the APCER and BPCER values show the algorithm is still inconsistent in its classification. The calculations indicate that the model cannot be used for any meaningful classification of bona fide or morphed images without more optimization and training.

We discovered using LRP that our fine-tuned neural network, with the highest accuracy, concentrates on the eyes, nose, and mouth areas while detecting morphed facial images. For the most part, the rest of the image is ignored. While focusing on these features may be adequate to achieve relatively high accuracy, it has limitations. There is a high degree of inconsistency in what the algorithm deems relevant between the different ways of visualizing the input relevance. This could have unforeseen problems, which is especially serious for security-related applications. In critical systems, the algorithm should consider data from all image locations during the classification process to achieve robustness.

Due to the complexity of the behaviour of a CNN's fully connected layers, the relevance scores generated by LRP are not immediately interpretable, demanding additional research to comprehend the network's overall behaviour completely. Seibold et al. [SHE21b] found that LRP commonly assigns high relevance scores to artefact-free regions in morphed face shots, implying that these regions are critical for the decision to classify the images as morphs. This aligns with our results where we see the neural network assign relevance

to areas without any visible artefacts and fails to detect other areas with clearly visible artefacts. In Figure 2 Bottom, this is visible where the hairline of multiple morphed images has artefacts that the algorithm does not detect very well.

Using LRP to visualize the decision-making process of convolutional neural networks fine-tuned for morph attack detection has been shown to be inconsistent. Problems with limited high-quality large datasets are an issue that makes the results of LRP challenging to assess and necessitates further research.

6 Conclusion

Given the strong generalization capabilities of face recognition systems, an adversary can execute targeted attacks that use morphed face images, as presented by Ferrara et al. [FFM14]. Face morphing attacks are a significant and dangerous concern, given that several countries enable residents to provide a picture for passports or national identification cards. Without requiring specialist knowledge, these photos can be forged utilizing readily available tools or websites [Se20]. Advances in deep learning and machine learning approaches have enabled the development of relatively good-quality morphs through various novel techniques. Generalizing morph attack detection is still a long way off, given the fundamental difficulty of gathering massive public databases with a variety of morph production strategies [Ve21]. Robust MAD algorithms must account for the vast diversity of picture post-processing, printing, and scanning technologies. The observed accuracy for detecting face image morphing attacks does not yet represent generalization to datasets containing a range of real-world capture situations [Sc19].

We discovered through LRP that a fine-tuned neural network focuses mainly on the eyes, nose, and mouth to detect morphed images. Though neural network analysis is still in its infancy, we demonstrated how methods such as LRP may be utilized to get valuable insights into a neural network's decision-making process. Future strategies for modifying the training process, particularly the training data, to increase resilience can be developed from this knowledge. Through the experiments, we got insights into how to train a network specifically for detecting face morphs and, more broadly, visual results on what the neural network used in its decision-making. High accuracy does not always imply robustness, implying the necessity for additional quality measures. We presented relatively high training and validation accuracy metrics for our MAD algorithm. LRP showed that most of the input images got ignored, implying a lack of robustness. In addition, our results show inconsistency in the algorithm's ability to detect artefacts and other features of morphed images. Future work could improve the used dataset and codebase, trying to decrease the computational cost of training the neural network while still being able to calculate essential performance metrics for the MAD algorithm.

References

[Ba15] Bach, Sebastian; Binder, Alexander; Montavon, Grégoire; Klauschen, Frederick; Müller, Klaus-Robert; Samek, Wojciech: On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. PloS one, 10(7):e0130140, July 2015.

[FFM14] Ferrara, Matteo; Franco, Annalisa; Maltoni, Davide: The magic passport. In: IEEE International Joint Conference on Biometrics. p. pp. 4, 2014.

[FFM19] Ferrara, Matteo; Franco, Annalisa; Maltoni, Davide: Decoupling texture blending and shape warping in face morphing. In: 2019 International Conference of the Biometrics Special Interest Group (BIOSIG). pp. 1–5, 2019.

[IS17] ISO/IEC DIS 30107-3, Information technology - Biometric presentation attack detection - Part 3: Testing and reporting.

[La16] Lapuschkin, Sebastian; Binder, Alexander; Montavon, Grégoire; Müller, Klaus-Robert; Samek, Wojciech: The LRP toolbox for artificial neural networks. The Journal of Machine Learning Research, 17(1):pp. 1–4, 2016.

[Ng20] Ngan, M.; Grother, Patrick J.; Hanaoka, Kayee K.; Kuo, J.: , "Face Recognition Vendor Test (FRVT) part 4: MORPH - Performance of Automated Face Morph Detection"National Institute of Technology (NIST), 2020.

[Ra17] Raghavendra, R.; Raja, KiranB.; Venkatesh, Sushma; Busch, Christoph: Face morphing versus face averaging: Vulnerability and detection. In: 2017 IEEE International Joint Conference on Biometrics (IJCB). pp. 555–563, 2017.

[Ra21] Raulf, Arne P; Däubener, Sina; Hack, Ben; Mosig, Axel; Fischer, Asja: SmoothLRP: Smoothing LRP by Averaging over Stochastic Input Variations. In: ESANN. 2021.

[Sc19] Scherhag, Ulrich; Rathgeb, Christian; Merkle, Johannes; Breithaupt, Ralph; Busch, Christoph: Face Recognition Systems Under Morphing Attacks: A Survey. IEEE Access, 7:pp. 23012–23014, 23016, 23019–23026, 2019.

[Sc20] Scherhag, Ulrich; Rathgeb, Christian; Merkle, Johannes; Busch, Christoph: Deep Face Representations for Differential Morphing Attack Detection. IEEE Transactions on Information Forensics and Security, 15:pp. 3625–3637, 2020.

[Se20] Seibold, Clemens; Samek, Wojciech; Hilsmann, Anna; Eisert, Peter: Accurate and robust neural networks for face morphing attack detection. Journal of Information Security and Applications, 53:pp. 1–6, 8, 10–11, 2020.

[SHE21a] Seibold, Clemens; Hilsmann, Anna; Eisert, Peter: Feature Focus: Towards Explainable and Transparent Deep Face Morphing Attack Detectors. Computers, 10(9):pp. 1–7, 9, 12, 14, 2021.

[SHE21b] Seibold, Clemens; Hilsmann, Anna; Eisert, Peter: Focused LRP: Explainable AI for Face Morphing Attack Detection. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) Workshops. pp. pp. 88–92, 95, January 2021.

[TB21] Tapia, Juan E.; Busch, Christoph: Single Morphing Attack Detection Using Feature Selection and Visualization Based on Mutual Information. IEEE Access, 9:pp. 1–2, 5–6, 10, 2021.

[Ve20] Venkatesh, Sushma; Zhang, Haoyu; Ramachandra, Raghavendra; Raja, Kiran; Damer, Naser; Busch, Christoph: Can GAN Generated Morphs Threaten Face Recognition Systems Equally as Landmark Based Morphs? - Vulnerability and Detection. In: 2020 8th International Workshop on Biometrics and Forensics (IWBF). pp. 1–6, 2020.

[Ve21] Venkatesh, Sushma; Ramachandra, Raghavendra; Raja, Kiran; Busch, Christoph: Face Morphing Attack Generation and Detection: A Comprehensive Survey. IEEE Transactions on Technology and Society, 2(3):pp. 128–145, 2021.

[Zh16] Zhang, Kaipeng; Zhang, Zhanpeng; Li, Zhifeng; Qiao, Yu: Joint Face Detection and Alignment Using Multitask Cascaded Convolutional Networks. IEEE Signal Processing Letters, 23(10):1499–1503, 2016.

[Zh21] Zhang, Haoyu; Venkatesh, Sushma; Ramachandra, Raghavendra; Raja, Kiran; Damer, Naser; Busch, Christoph: MIPGAN—Generating Strong and High Quality Morphing Attacks Using Identity Prior Driven GAN. IEEE Transactions on Biometrics, Behavior, and Identity Science, 3(3):365–383, July 2021.