Explainable Deep-Fake Detection Using Visual Interpretability Methods

Badhrinarayan Malolan, Ankit Parekh, Faruk Kazi
Centre of Excellence (CoE) in Complex and Non-linear Dynamical Systems (CNDS),
Veermata Jijabai Technological Institute
Mumbai, India
e-mail: badhrinarayan_b16@et.vjti.ac.in, ajparekh_b16@et.vjti.ac.in, fskazi@el.vjti.ac.in

***Abstract*—Deep-Fakes have sparked concerns throughout the world because of their potentially explosive consequences. A dystopian future where all forms of digital media are potentially compromised and public trust in Government is scarce doesn't seem far off. If not dealt with the requisite seriousness, the situation could easily spiral out of control. Current methods of Deep-Fake detection aim to accurately solve the issue at hand but may fail to convince a lay-person of its reliability and thus, lack the trust of the general public. Since the fundamental issue revolves around earning the trust of human agents, the construction of interpretable and also easily explainable models is imperative. We propose a framework to detect these Deep-Fake videos using a Deep Learning Approach: we have trained a Convolutional Neural Network architecture on a database of extracted faces from FaceForensics' DeepFakeDetection Dataset. Furthermore, we have tested the model on various Explainable AI techniques such as LRP and LIME to provide crisp visualizations of the salient regions of the image focused on by the model. The prospective and elusive goal is to localize the facial manipulations caused by Faceswaps. We hope to use this approach to build trust between AI and Human agents and to demonstrate the applicability of XAI in various real-life scenarios.**

Keywords-deep-fakes; deep-fake detection; faceswap; interpretability; explainable AI (XAI); LRP; LIME

I. INTRODUCTION

The onset of Deep-Fakes has been marked by Deep Generative Algorithms such as Generative Adversarial Networks (GANs) and Convolutional Autoencoders trying to outdo each other to create the perfect Deep-Fake video. The rapid advancement of this highly sophisticated and novel technology has left the world in awe and apprehension simultaneously. No longer are people constrained by the lack of sufficiently advanced hardware, as popular applications such as FakeApp can easily produce convincing Faceswap videos. Faceswaps, in particular, have created a massive buzz in social media due to their novelty and the possibly damaging effects it can have on society, such as defamation and blackmail. Due to the massive potential of misuse and misinformation held by these forged videos, it becomes necessary to create accurate and robust models to detect these fake media.

Our research aims to develop a pipeline for the detection of these Deep-Fake videos. The main focus however, will not be maximizing the accuracy of our model on a single or combination of datasets. Instead, we aim to create simple and easily understandable visual interpretations of our model for a given set of input images. The black-box approaches imposed on us by largely opaque Deep Learning techniques have alienated applications where model interpretability is a primary concern, e.g.: Biomedical applications, Healthcare, Financial domains including High-Frequency Trading, etc.

We have been inspired by the rise of the field of "Explainable Artificial Intelligence" (XAI) which aims to demystify the various approaches of Machine Learning and Deep Learning, and allows the internal working of these models to be more transparent, providing easy-to-explain interpretations of their decisions to a human audience. The term XAI was popularized by DARPA's XAI program [1]. The roots of this term, however, can be traced to [2] and the pursuit of model interpretability dates even further back to the 90s with [3] using Saliency Maps for the interpretation of Neural Networks. There are significant benefits to investing in XAI, including business returns by means of satisfying investors and compliance with GDPR legislation, a part of which includes "Right to Explanation" currently in effect across the EU. More importantly, it caters to the fundamental social Responsibility of openness and ethical behaviour.

Deep-Fakes have spawned a whole new sub-area of Deep-Fake Detection methods which aim to discriminate between genuine videos and forged videos. Different approaches deal with the problem by either using Deep-Fake videos as they are i.e. exploiting the temporal nature of the video or by extracting frames from the video. D. Afchar et al. [4] have analyzed the mesoscopic qualities of Deep-Fake images from Faceswap and Face2Face methods of manipulation using two Convolutional Neural Network (CNN) architectures namely: Meso-4 and MesoInception-4 to perform binary classification. D. Guera et al. [5] implements a Convolutional-LSTM network, the former of which extracts frame-level features and the latter does Sequence Processing fed into Dense layers which ultimately determine if the video is real or forged. Deep-Fakes, especially Faceswaps are never perfectly composed, with noticeable warping and blurring around the face area being present in some frames which usually give them away. Li et al. [6] have taken advantage of this by capitalizing on the resolution mismatch of swapped faces and the presence of artifacts due to affine transforms and trained a CNN to capture these features. The FaceForensics++ dataset introduced by A. Rossler et al. [7] is perhaps one of the more important contributions to the challenge of Deep-Fake Detection with the curation of a pristine Deep-Fake dataset

978-1-7281-7283-5/20/$31.00 ©2020 IEEE
DOI 10.1109/ICICT50521.2020.00051

and introduction of a Benchmark for performance on the same. But the question comprehending model behaviour remains unanswered.

Our work would be incomplete without an extensive perusal of the landscape of XAI and the various methods of analyses of black box Deep Learning models to give us crisp visual representations superimposed on our input image. To achieve our goal of the Explainability of Deep-Fake detection, we continue with a survey of some XAI methods.

The introduction of saliency maps in K. Simonyan et al. [8] as an analysis tool piqued the interest of researchers to look more closely into their CNN. Selvaraju et al. [9] showcased Grad-CAM, a method to localize regions in the image that are important for its predicted class, with further applications in VQA (Visual Question Answering), which is recommended for interpretation of CNNs for image classification.

Ribeiro et al. [10] introduced LIME, which preserves local fidelity to localize the interpretation of the model around the instance predicted. LIME has proved to be a versatile method to generate explanations from different kinds of Machine Learning models regardless of the model architecture since it doesn't need to glance into the model itself.

One of the most powerful and effective XAI methods out there is Layer-Wise Relevance Propagation (LRP) introduced by Bach et al. [11]. It evaluates a relevance score for every neuron by doing a backward pass in the deep neural network, thus elucidating why a specific decision was taken. LRP and its variants feature extensively in the results of this paper to plot visual heatmaps and highlight the salient features of the images.

The paper is organized as follows. Section II describes preliminaries and defines important key terms and concepts referred to throughout the paper. In Section III we finally expand on the methodology we used to detect Deep-Fake images and Section IV provides a sharp look into our results. The final section is reserved to provide a summary of our work and closing statements.

II. Preliminaries

To provide context, we have shortly summarized the relevant terms in this section. We have also explained some of the interpretability methods that we have experimented with, the results of which we will display in subsequent sections.

A. Deep-Fakes

Deep-Fakes, in general, can be described as fake media (video, images, audio, text) generated by Deep Learning Algorithms such as GANs and Autoencoders. The most common form of video Deep-Fakes are Faceswaps, where a pair of encoders trained on a dataset of two target faces, condense the features of their respective faces into a lower-dimensional embedding and a decoder subsequently transforms each of the faces into the other by using the other face's embedding. The impressive level of realism achieved by this technique, the large volume of freely available training data and increasing access to advanced computing resources such as GPUs have resulted in the mass proliferation of these videos.

B. Dataset

The dataset [12] we are using is a subcomponent of the popular FaceForensics++ Dataset of The Technical University of Munich namely, the DeepFakeDetection (Google) Dataset which was curated by Google and Jigsaw. It consists of 363 original source actor videos as the real counterpart and 3068 manipulated videos as the fake counterpart created by shooting videos of paid consenting actors and faceswapping them.

C. Model Architecture

We will use the Xception network introduced by F. Chollet [13], which is a traditional CNN with Depth-wise Separable Convolutions. This network was designed with the express purpose of outperforming the traditional Inception architecture on the ImageNet Database i.e. image classification tasks. Hence, this network will help us extract powerful distinguishing features from our images.

D. Local Interpretable Model-Agnostic Explanations (LIME)

LIME is a technique to interpret the predictions of any classifier with easy to comprehend explanations. It preserves local fidelity around the specific instances of predictions given by the model, i.e. it ensures the local behavior of the model around a particular prediction instance.

We denote the classification model under consideration as follows: $f: \mathbb{R}_d \rightarrow \mathbb{R}$ and $g \in G$ as the explanation of the said model. Hence, the classification probability of a particular input x would be $f(x)$. This probability acts as a binary signal for the association of x with a certain class. To define a locality measure around x we use $\pi_x(z)$ to represent proximity between an instance z and x. To measure the unfaithfulness of g we define a loss measure $\mathcal{L}(f, g, \pi_x)$ for the locality of π_x. Finally, we minimize this loss $\mathcal{L}$ while keeping the complexity of the explanation $\Omega(g)$ low enough to preserve local fidelity and ensure interpretability by humans.

Following is the explanation provided by LIME:

$$\xi(x) = \underset{g \in G}{argmin}\ \mathcal{L}(f, g, \pi_x) \ + \ \Omega(g) \qquad (1)$$

The Loss Function is defined as:

$$\mathcal{L}(f, g, \pi_x) = \sum_{z, z' \in Z} \pi_x(z)(f(z) - g(z'))^2 \qquad (2)$$

Here, z is a perturbed data point in the original data space, z' is the corresponding interpretable representation and $\pi_x(z)$ weighs the samples based on its similarity to data point x. Using the LIME [14], we present intepretable visualizations of our input image superimposed on the prediction based attention slice of our model.

E. Layer-Wise Relevance Propagation (LRP)

Traditional Sensitivity Map and Saliency methods reveal little about the function whose output they represent. Layer-wise relevance propagation (LRP) operates by building a local redistribution rule for each neuron of a deep network and produces a pixel-wise decomposition by applying these rules in a backward pass.

$$z_j = \sum_i z_i \times w_{ij} + b_j \quad (3)$$

Consider a deep neural network consisting of layers of neurons, the output of an upper-layer single neuron z_j would be, where z_i are the outputs of lower-layer neurons in the forward pass and w_{ij} , b_j are weights and biases of corresponding connections between neurons.

At first, the relevance score of the output layer neurons is initialized to the prediction score of the target class of interest c , $f_c(x)$, followed by the computation of a layer-by-layer relevance score for lower level neurons. This is done by computing a relevance message $R_{i\leftarrow j}$ from a lower level neuron to all its children, upper level neurons.

$$R_{i\leftarrow j} = \frac{z_i \times w_{ij} + \frac{\epsilon \times sign(z_j) + \delta \times b_j}{N}}{z_j + \epsilon \times sign(z_j)} \times R_j \quad (4)$$

Then all these relevance messages are summed up to compute the relevance score of that lower level neuron. This process is continued till the input layer neurons.

$$R_i = \sum_j R_{i\leftarrow j} \quad (5)$$

III. Deep-Fake Detection

A. Extraction of Faces

There was an imbalance due to a large number of combinations of the real videos to generate fake sequences. To remedy this issue, frames were extracted from real and fake video sequences at different sampling rates to balance the real and fake classes of the dataset. We created a face extraction pipeline to create the dataset of images comprising of fake and real classes with train, test and validation splits of approximately 4:1:1. Refer Figure 1 for a pictorial description of our pipeline:

Figure 1. Face extraction pipeline.

We used the Dlib face extractor which identifies 68 facial landmarks on our image. Faces were extracted from the frames by choosing their central landmark and then cropping out a square that included different extents of background which aided us to monitor the focus area of the network in the later stages. We have trained our model on datasets of images with two different scales of background namely 1.3x and 2x with the faces occupying roughly 80 to 85% and 60 to 65% area respectively. Frames with no faces or with a number of faces more than one were rejected as in case of fake sequences, it was not possible to check which face was forged before training the model. The images (faces with background) had a range of dimensions so we resized all of them to 128 x 128 x 3 and rejected the ones that were initially found to be below 120 x 120 x 3. The final statistics of our dataset are provided in Table I, where there are ~24% more fake images than real ones.

TABLE I. Dataset Details

Dataset	Real Images	Fake Images
Train	67203	83728
Val	14696	17488
Test	14929	17513

B. Training

We implemented the Xception network comprising of 134 layers and a total of 20,863,529 parameters out of which 20,809,001 were trainable. We modified the input layer to accept a 128 x 128 x 3 image and changed the final activation to a sigmoid layer for our purposes of binary classification. We used the binary cross-entropy loss function over MSE loss to better facilitate convergence to the global minima.

The loss function was used in conjunction with Adam adaptive learning rate optimizer which works well even with little tuning of hyperparameters. We used no pretraining in our Network and trained the model entirely from scratch to visualize the raw features learned by the model on its own. We used the Keras Machine Learning Library on a Tensorflow backend to carry out all our experiments and trained the model for 20 epochs.

IV. Results

TABLE II. Summary of Results

Image Scale	Test Accuracy
1.3x	94.33%
2.0x	90.17%

In this section, we analyse the results obtained in Table II where test accuracies of both scales are given. In case of the 1.3x model, we have not fed any significant background data to the model apart from the face itself. Hence, we can expect the model to accurately identify the features of the face, and unsurprisingly it fares better in the test set compared to the 2x model. We reiterate that our focus is not to achieve the highest accuracy on a benchmark.

Considering this, we decided to showcase the results of only the 2x model. It would be interesting to see how well the model performs with a lot of background information. Based on the theme of our topic, we'll be showing only fake faces throughout this section. Also, we will test the robustness of our model to Gaussian blur noise and affine transforms. generate a better oversampled dataset that can be fed to models to learn the entire context of data.

A. Intermediate Activations

We have showcased the activations of the block2_sepconv2_act layer for an input image in Figure 2. This particular layer detects various forms of edges present

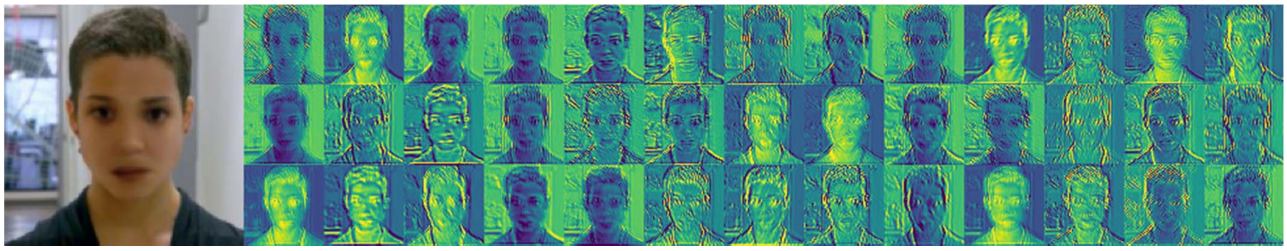

Figure 2. Input image and a slice of block2_sepconv2_act (Activation of the second sepconv layer of the second block).

throughout the human face like the forehead, eyes, mouth, and jawline, while the activation retains most of the information present. The activations of the deeper layers become increasingly abstruse and less visually interpretable, as they start extracting complex features like shapes of eyes, nose, and ears which are often deciding factors in localizing manipulations. The filters present in the deepest layers learn the most complex attributes throughout the network. Hence, they mostly go unactivated because these features are usually not present in our input. Hence, our network effectively decomposes large features into smaller and more complex attributes with raw data (faces with background) getting transformed filtering background information while useful facial information gets magnified and refined.

B. *Local Interpretable Model-Agnostic Explanations (LIME)*

Figure 3. 1st Row shows the input along with its perturbations and 2nd row shows the LIME descriptions of these inputs.

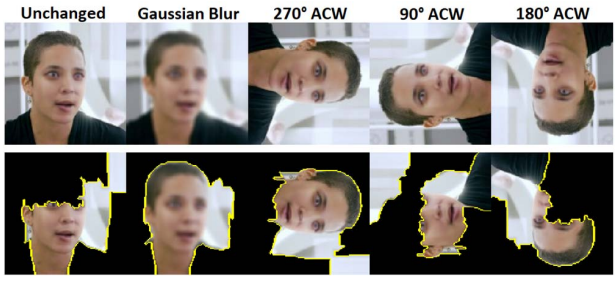

Figure 4. 1st Row shows the input along with its perturbations and 2nd row shows the LIME descriptions of these inputs.

From Figure 3, we see that LIME is able to capture the relevant areas surrounding the face that resulted in its classification. Little background data occupies the relevant slice, hence we can visually confirm that the model is looking at the right place.

The outputs seen in Figure 4 show that LIME responds well to affine transformations as well as Gaussian blur noise, despite not being trained on those images. Hence, we conclude that LIME has produced results favourable for our causes.

C. *Layer-Wise Relevance Propagation (LRP)*

With the help of iNNvestigate [15], we have managed to leverage their suite of LRP methods to sufficiently localize the regions of manipulation to the nose and mouth region in our input image as shown in Figure 6 on the next page. All the given methods are reformulations of the fundamental principle of LRP with each relevance score of each neuron being backpropagated all the way to the input layer, LRP Sequential being designed specifically for CNNs. The heatmap suggests that the artifacts found in the nose and mouth region of the face are essential for the classification of this particular image.

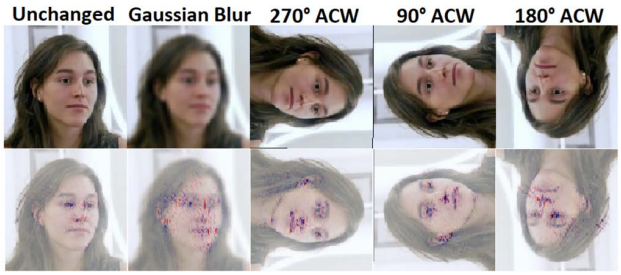

Figure 5. `lrp.sequential_preset_b_flat` compared on various input perturbations.

We've also displayed the results of input perturbations on LRP in Figure 5. We have fixed the method used as `lrp.sequential_preset_b_flat` and applied the same variations (Gaussian Blur and affine transforms) to the input and contrasted it with the input heatmap. We see that LRP has preserved the structural qualities of the original heatmap, albeit being a bit noiser. These experiments aim to firmly establish the validity of these methods and contrast its performance on a variety of inputs.

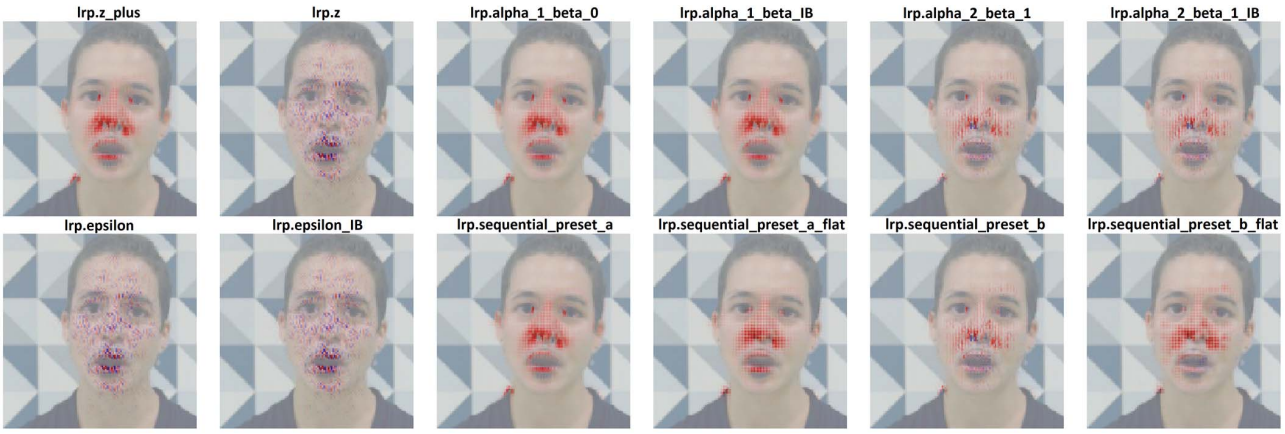

Figure 6. Comparison of different heatmaps produced by the different LRP rules.

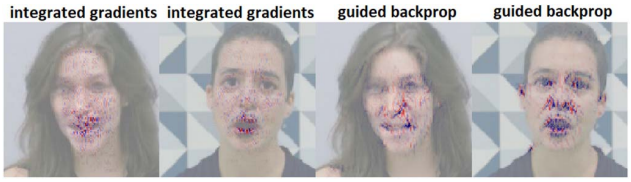

Figure 7. We have obtained favourable results on two more methods namely Integrated Gradients and Guided. Backpropagation however, they are not the main methods of focus of our paper.

V. Conclusion

We have obtained a collection of results that propose explanations to the predictions given by our classifier model in terms of heatmaps or image concept slices, also the input perturbation results point to our model achieving rotational invariance to a large extent. Hence, we have substantiated the performance of our model towards the task of detecting Deep-Fake images from a video in a way that even a lay-person can be convinced. Striking similarities are observed between these models in terms of the regions of interest highlighted by them, with many of them focusing their attention on the same regions in the image. In this way, the utilization of XAI techniques furthers our understanding of complex models and provides an arena to present some much-needed context to the seemingly obtuse decisions arrived at by AI. We hope that, as a result of this research, the end goal of cultivating trust between AI practitioners and the target customers is a little closer.

Acknowledgment

This work was supported in part by the Centre of Excellence in Complex and Nonlinear Dynamical Systems (CoE-CNDS) and in part by the Veermata Jijabai Technological Institute (VJTI), Matunga, Mumbai, India, under the Technical Education Quality Improvement Programme (TEQIP-III, subcomponent 1.2.1).

References

[1] D. Gunning DARPA Explainable Artificial Intelligence (XAI) Program. http://www.darpa.mil/program/explainable-artificial-intelligence 2016. [Online; accessed 10-November-2019]

[2] M. Lent, W. Fisher and M. Mancuso (2004). An Explainable Artificial Intelligence System for Small-unit Tactical Behavior. In Proceedings of the Nineteenth National Conference on Artificial Intelligence, 16th Conference on Innovative Applications of Artificial Intelligence.

[3] N.J.S. Morch, U. Kjems, L.K. Hansen, C. Svarer, I. Law, B. Lautrup, S. Strother and K. Rehm (1995). Visualization of neural networks using saliency maps. In Proceedings of ICNN'95 - International Conference on Neural Networks.

[4] D. Afchar, V. Nozick, J. Yamagishi and I. Echizen (2018). MesoNet: a Compact Facial Video Forgery Detection Network. arXiv:1809.00888v1, 2018

[5] D. G¨uera and E. J. Delp (2018). Deepfake Video Detection Using Recurrent Neural Networks. In IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS).

[6] Y. Li and S. Lyu (2019). Exposing DeepFake Videos By Detecting Face Warping Artifacts. arXiv:1811.00656v3, 2019.

[7] A. Rossler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies and M. Nießner (2019). FaceForensics++: Learning to Detect Manipulated Facial Images arXiv:1901.08971v3, 2019.

[8] K. Simonyan, A. Vedaldi and A. Zisserman (2014). Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps. arXiv:1312.6034v2, 201

[9] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra (2017). Grad-cam: Visual explanations from deep networks via gradientbased localization. In Proceedings of the IEEE International Conference on Computer Vision.

[10] M. Ribeiro, M. Tulio, S. Singh, and C. Guestrin. "Why Should I Trust You?: Explaining the predictions of any classifier." Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. ACM, 2016.

[11] S. Bach, A. Binder, G. Montavon, F. Klauschen, K. Muller, and W. Samek. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. PLOS ONE, 10(7):e0130140, 2015.

[12] Dataset available via https://github.com/ondyari/FaceForensics

[13] F. Chollet. Xception: Deep Learning with Depthwise Separable Convolutions. In IEEE Conference on Computer Vision and Pattern Recognition, 2017.

[14] LIME: https://github.com/marcotcr/lime

[15] iNNvestigate: https://github.com/albermax/innvestigate