Towards Explainability in Using Deep Learning for Face Detection in Paintings

Siwar Bengamra, Olfa Mzoughi, André Bigand, Ezzeddine Zagrouba

Towards Explainability in Using Deep Learning for Face Detection in Paintings

Siwar Bengamra[1,3][a], Olfa Mzoughi[2][b], André Bigand[3][c] and Ezzeddine Zagrouba[1][d]
[1]*LIMTIC Laboratory, Higher Institute of Computer Science, University of Tunis El-Manar, Ariana, Tunisia*
[2]*Prince Sattam Bin Abdulaziz University, Kingdom of Saudi Arabia, Saudi Arabia*
[3]*LISIC Laboratory, University of the Littoral Opal Coast (ULCO), Calais Cedex, France*

Keywords: Explainable Artificial Intelligence (XAI), Perturbation-Based Explanation Method, Object Detection, Painting Images.

Abstract: Explainable Artificial Intelligence (XAI) is an active research area to interpret a neural network's decision by ensuring transparency and trust in the task-specified learned models. In fact, despite the great success of deep learning networks in many fields, their adoption by practitioners presents some limits, one significant of them is the complex nature of these networks which prevents human comprehension of the decision-making process. This is especially the case in artworks analysis. To address this issue, we explore Detector Randomized Input Sampling for Explanation (DRISE), a visualization method for explainable artificial intelligence to comprehend and improve CNN-based face detector on Tenebrism painting images. The results obtained show local explanations for model's prediction and consequently offer insights into the model's decision-making. This paper will be of great help to researchers as a future support for explainability of object detection in other domain application.

1 INTRODUCTION

Deep learning models have shown great success in various computer vision applications such as image classification, semantic analysis and object detection. However, alongside their impressive performances and their increasing applicability, the deep neural networks are perceived as "blackbox" model and difficult to interpret and to explain to the end users due to its high non-linear computations. Thus, by debugging and auditing the model, explainability can detect faulty behavior important for performance improvement. The poor explainability could also decrease human trust in the model with a risk of abandoning its use, especially in critical applications such as medical imaging since clinicians confidence is necessary for adoption (Tulio Ribeiro et al., 2016). The demand of eXplainable Artificial Intelligence (XAI) is becoming increasingly essential for model prediction as the deep neural networks become the algorithm of choice for models. A popular approach for explanation of the deep learning model is the use of attribution principle which aims at computing scores over pixels of the input image, reflecting the importance of each pixel to the output of the model (Petsiuk et al., 2021). The arrangement of the ranking of all input pixels forms a heatmap also called pixel attribution map, saliency map, or more generally an explanation map.

In particular, the object detection in Tenebrism paintings presents considerable challenges according to large variation of contrast, limited color palette and limited number of available images (Mzoughi et al., 2018; Gamra et al., 2021). The great success that deep Convolutional Neural Networks (CNNs) have achieved for object detection in Tenebrism images (Gamra et al., 2021) raises fundamental questions on the internal working. For example, how the hidden layers process this kind of images? Is there a decisive hidden layer or stage (i.e. serie of hidden layers) in the network's architecture? What were the features that contributed to the model's output for a given such input (i.e. Tenebrism painting). Is all subregions within the object's bounding box are equally important for the model decision? These questions become more important as the importance of object detection in Tenebrism images based deep learning increases. Interestingly, in this art painting domain,

[a] https://orcid.org/0000-0001-5546-5292
[b] https://orcid.org/0000-0001-8758-9740
[c] https://orcid.org/0000-0002-3165-5363
[d] https://orcid.org/0000-0002-2574-9080

Bengamra, S., Mzoughi, O., Bigand, A. and Zagrouba, E.
Towards Explainability in Using Deep Learning for Face Detection in Paintings.
DOI: 10.5220/0011670300003411
In *Proceedings of the 12th International Conference on Pattern Recognition Applications and Methods (ICPRAM 2023)*, pages 832-841
ISBN: 978-989-758-626-2; ISSN: 2184-4313

explainability methods can bring benefits both to the model performances improvement and user trust increase. Firstly, this can help the researchers to analyze and justify object detection results by providing the required information about feature extraction. An enhanced control can also be ensured by identifying and correcting unexpected behaviour. Finally, explaining and understanding the internal mechanics could offer possibilities to improve the model. For example, understanding outliers or missing values allows users know how to make the model smarter.

In this paper, we focus on understanding how face detector models based deep learning work and producing insights into model's decision process such as source of failures. So we investigate an implementation of DRISE (Petsiuk et al., 2021) to generate visual explanation for our face detection results, then use this for models comparison and failure's source investigation. The choice of this method was motivated by adopting the attribution approach based perturbation. From the few previous perturbation-based methods intended for object detectors, we find the D-RISE (Petsiuk et al., 2021) method particularly interesting since it is based on pixel-wise perturbation promising to generate saliency maps more accurately in terms of location. The rest of the paper is organized as follows. Section 2 briefly review the existing works employing explainability in painting images. Then, section 3 reviews and discusses the literature on explainability methods. A detailed description of the DRISE method will be present in section 4. Section 5 provides experimental results and analysis. Concluding remarks and potential future research directions are presented in section 6.

2 RELATED WORKS

Despite the considerable amount of works and surveys devoted to generate explanations for deep CNNs, little attention has been paid to explain the model's predictions in paintings to human users (see table 1).

This section surveys the related previous research in the field of explainability deep CNNs used for artwork analysis tasks. In (Cetinic et al., 2019), the authors investigated an attention mechanism in order to highlight the regions responsible for predicting the aesthetic, sentiment, and memorability scores in the context of art history. The obtained attribution maps were generated by computing probability weight for each image location based on image features, hidden layers and softmax function. The obtained attribution maps was a good yardstick to compare the used CNN models. In a recent work (Surapaneni et al., 2020), the Gradient Weighted Class Activation Maps (GradCAM) method (Selvaraju et al., 2017) has shown its effectiveness to add transparency and explainability into a deep learning model used for classifying artworks. Earlier, in (Pinciroli Vago et al., 2021), the authors used Class Activation Maps (CAM) method (Zhou et al., 2016) to explain how the classification of characters in Christian art paintings works by localizing areas of a painting contributing the most to the output result. Several CAM variants such as GradCAM++ (Chattopadhay et al., 2018) and Smooth GradCAM++ (Omeiza et al., 2019) are compared in terms of their capacity to identify the iconographic features required for the classification task.

3 STATE OF THE ART EXPLAINABILITY METHODS

Two broad categories of attribution-based explanation methods exist, namely grdient-based methods and perturbation-based methods (see figure 1).

3.1 Gradient-Based Methods

Gradient-based methods (or Backpropagation-based methods) compute the attribution scores by calculating the gradients of the model's output with respect to the extracted features or input via back-propagation algorithm. The paper (Simonyan et al., 2014) provides a simple method for generating saliency map by differentiating the output of the model with respect to the input. Later, several methods (Springenberg et al., 2014; Zeiler and Fergus, 2014; Smilkov et al., 2017) have been proposed to enhance the saliency maps by reducing the visual noise. Some trend to modify the gradients of ReLU functions by removing negative values during the back-propagation computation, while others average the gradient over multiple inputs with additional noise. Particularly, Class Activation Mapping based methods (Zhou et al., 2016), abbreviated as CAM, perform Global Average Pooling on the last feature map and pass the pooled features to the fully connected layer. Then, the predicted class score is mapped back to the previous convolutional layer to generate the importance maps. The application of the (Zhou et al., 2016) method has been limited to specific CNN architectures trained with a Global Average Pooling (GAP) layer injected between the last convolutional layer and the final fully connected layer. Several more sophisticated methods have subsequently been proposed. For example, GradCAM (Selvaraju et al., 2017) was a generalization of CAM that can generate visualizations for any classification CNN, re-

Table 1: Summary of papers focusing on explaining deep neural networks in painting.

Reference	Explainability method	Task	Artwork dataset	Evaluation method
(Pinciroli Vago et al., 2021)	CAM, GradCAM, GradCAM++, and Smooth GradCAM++	Classification of iconographic elements	Christian art paintings	- Qualitative analysis - Quantitative analysis: Intersection Over Union, Bounding box coverage, Irrelevant attention
(Surapaneni et al., 2020)	GradCAM	Image classification	Paintings from the Met's online collection	Qualitative evaluation
(Cetinic et al., 2019)	Soft attention mechanism	Predicting aesthetic, sentiment, and memorability scores	Paintings from WikiArt collection- fine Art	Qualitative evaluation

gardless of its architecture. GradCAM++ (Chattopadhay et al., 2018) presents also an extension of the GradCAM that can generate improved visual explanations using a weighted combination of the positive partial derivatives. Other extensions of CAM proposed the modification of back-propagation rules to have a probabilistic or local approximation backpropagation scheme. For example, in Excitation Backprop (Zhang et al., 2016), authors use stochastic sampling process to integrate the forward activations and backpropagated gradients efficiently.

3.2 Perturbation-Based Methods

Perturbation-based methods (or occlusion-based methods) compute the attribution of input pixels by perturbing their values (e.g. by occlusion, adding noise, blurring or modifying certain input pixels) and record the effect of these changes on the model performances (Ivanovs et al., 2021). This kind of methods performs explanation of the model viewed in terms of its inputs and outputs, without access to its internal functioning. Significant research on perturbation-based methods has been carried out. For example, (Zeiler and Fergus, 2014) proposed the occlusion method, which simply modify different contiguous rectangular patches of the input image with a given baseline (e.g. all zero patch) and evaluate the effect of this perturbation on the target output. Another explainability method is LIME (Local Interpretable Model agnostic Explanations) (Ribeiro et al., 2016) which divides the image into interpretable components (contiguous superpixels), generates a data set of perturbed images by masking some of the interpretable components and predicts the class probabilities. A linear model is then trained on this data set and the superpixel weights of that linear model are presented as an explanation of the prediction. Recently, Petsiuk et al. (Petsiuk et al., 2018) have proposed RISE (Randomized Input Sampling for Explanation of Black-box Models) as a method for classification decisions explanation. The RISE method element-wise multiply the input image with several random masks and pass the resulting perturbed images to the model. The saliency map is computed as a weighted sum of random masks, where weights are probability-like score for the masked images with respect to each class. More Recently, an extension of the RISE method (Petsiuk et al., 2018) to object detectors, called Detector Randomized Input Sampling for Explanation (D-RISE), is proposed in (Petsiuk et al., 2021).

3.3 Discussion

Although there are many interesting explainability methods, a discussion on the two categories of methods described previously is required to understand the differences and similarities between them. For the gradient-based methods, the main advantage is the computational speed, however, its drawbacks should also be considered for application. In fact, the gradient in discontinuities creates noisy saliency maps, especially in case of large deep neural networks (Ancona et al., 2019). Moreover, many backpropagation-based methods are limited to certain network architectures and/or layer types, and therefore are restricted in their use. For example, Guided backpropagation (GuidedBP) (Springenberg et al., 2014) is limited to CNN models with ReLU activation. GradCAM (Selvaraju et al., 2017) is also limited to specific architectures which use the AveragePooling layer to connect convolutional layers to fully connected layers. Additionally, it is still very challenging to determine the validity of gradient-based methods since they do not directly measure the effect of perturbing input images (Nielsen et al., 2022). Fo object detection explainability, it is not possible with gradient-based method to produce visual explanation for an arbitrary bounding box (i.e. not detected by the model) since there is no starting point to propagate from (Petsiuk et al., 2021).

Compared to the gradient methods, perturbation-based methods are totally independent of the model's

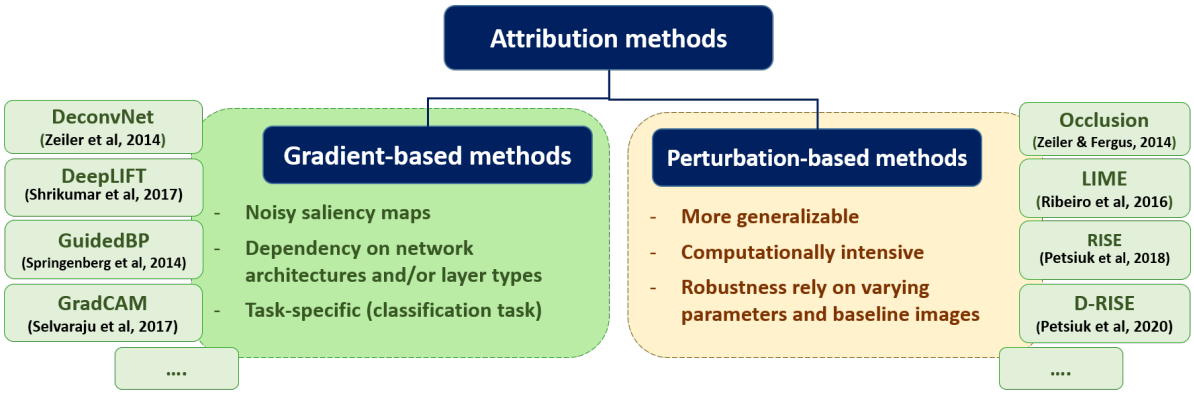


Figure 1: Categories of attribution methods.

architecture. An important advantage of perturbation-based methods is that do not require access to the internal parts of the model for explanation, and therefore are used to explain almost deep neural networks. Moreover, perturbation-based algorithms have become quite popular due to their effectiveness and ease of implementation (Qiu et al., 2021). On the other hand, it should pointed out that explanation based perturbation lacks consistency in the explanation. In fact, many choices of parameters lead to the divergence of saliency maps (Brunke et al., 2020; Thakur and Fischmeister, 2021). For example, the number of iterations used for explanation of the same image could result in different salient region detection. Attention should also be paid to the underestimation of the pixel importance values by not considering all perturbation directions (Kim et al., 2021). It is also to note that perturbation based methods tend to be very slow as the number of images pixels (or features) to test grows. Finally, a special attention should be given to the scope and shape of the perturbations which affect the granularity of the output saliency map (Ivanovs et al., 2021). Pixel-wise perturbation produce accurate saliency maps in terms of location, while patch-wise perturbation deliver maps where boundaries fit the object boundaries.

In conclusion, perturbation-based methods have several advantages over gradient-based methods, making them a popular promising approach in XAI. We remark also a lack of perturbation-based methods concerning the explainability of object detectors (Petsiuk et al., 2021; Padmanabhan, 2022; Hogan et al., 2022), while the main effort has been focused on the explainability of classification decisions.

4 DRISE EXPLAINABILITY METHOD

The Detector Randomized Input Sampling for Explanation (DRISE) method (Petsiuk et al., 2021) is a perturbation-based XAI method for object detectors. The process for generating the attribution maps was divided into five stages, namely, generating masks to be then applied to the input image, running object detector on masked images to get proposals, converting object detections (target and proposals) into vectors, computing similarities between target and proposals, and inferring saliency map. In the following, we discuss details of each stage of the DRISE pipeline applied on Tenebrism paintings, as shown in Figure 2. The process inputs are the input image **I**, the CNN object detector model **f** and the target object detection **T** specified by a bounding box (see green box in the figure 2) with a class probability. In our case, Faster RCNN is the detector network used to detect 'face' class. The objective is to explain a face detection result by generating a saliency map **S** highlighting pixels of the input image**I** considered important by the model **f** to get such prediction output **T**, namely the face's bounding box (green bounding box in Figure 2) and probability.

4.1 Mask Generation

As the first step, authors adopt the RISE masking technique (Petsiuk et al., 2018) to generate a set of random binary masks. This consists of sampling small binary masks and then upsampling them to larger resolution using bilinear interpolation. The input image I is element-wise multiplied with the obtained N masks (M_1.. M_N) to get N masked images (IM_1.. IM_N).

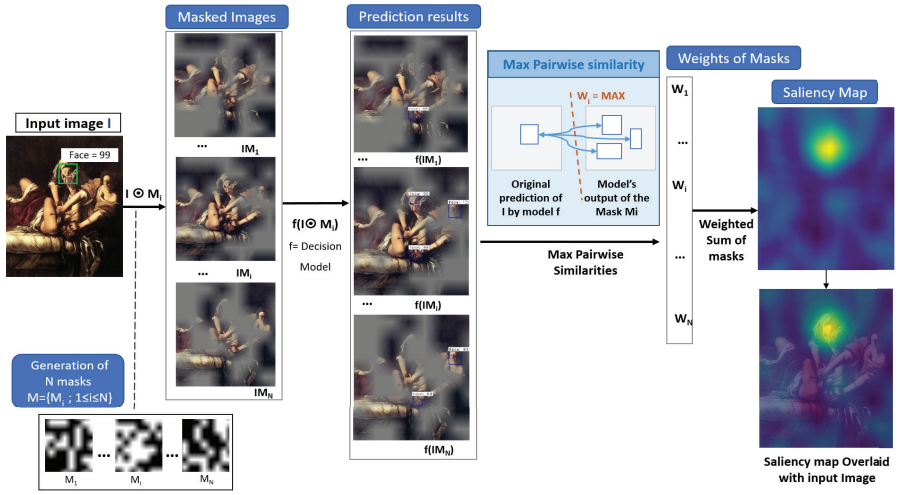


Figure 2: The pipeline of D-RISE method on Tenebrism painting.

4.2 Conversion of Detections

The detection result (target or proposal) is encoded into a vector containing the corners coordinates of the bounding box, the objectness score (0 if the bounding box contains an object of any class, 1 otherwise) and a one-hot vector for the probabilities $(p_1^i, ..., p_C^i)$ representing the probability that the bounding box belongs to each of the C classes. In our case, the faster RCNN based object detector does not produce objectness score and we have one class 'face'. So the vector consists of the bounding box corners and the probability to be a face.

4.3 Proposals Detection

This stage consists of running the object detector model **f** on the N masked versions of the input image to get the proposals (blue bounding boxes in Figure 2) that will be converted into detection vectors.

4.4 Similarity Computing

For each masked image, a pairwise similarities are computed between the target vector detection and all detection proposal vectors. Then the maximum of similarities is selected to be the weight of the mask in question.

4.5 Saliency Map Generation

The saliency map is computed as a weighted sum of the N masks. Note, that importance in **S** increases from blue colors to red ones.

5 EXPERIMENTS

In this section, we give an overview of the evaluation metrics, models, and used datasets, then we discusses the results of DRISE method on face detection results from painting images.

5.1 Experimental Settings

Dataset. We perform all our experiments on Tenebrism Dataset (Gamra et al., 2021).
CNN Models. In this work, we are exclusively considering deep learning models used for face detection in Tenebrism paintings called, Model1 and Model3 (Gamra et al., 2021). Specifically, the Model1 is a Faster RCNN based ResNet50, pretrained on AFLW Dataset (a famous dataset of photograph faces) and tested on the Tenebrism dataset. Concerning the Model3, it consists of transfer learning of Model1 by retraining all its layers on the target dataset (i.e Tenebrism dataset).
Attribution Method. We implemented the DRISE method to inject transparency and explainability into the desired CNN models. We follow the original setup. So the number of iteration and the parameters for mask generation are the same as in (Petsiuk et al., 2021).
Evaluation Metrics. We used the deletion and insertion metrics proposed by (Petsiuk et al., 2018) to evaluate the explanation maps. The deletion metric iteratively removes the N pixels most relevant to the face class by masking them with blurred ones, as illustrated in figure 3, measures its effect on the accuracy of face detection. On the other hand, the inser-

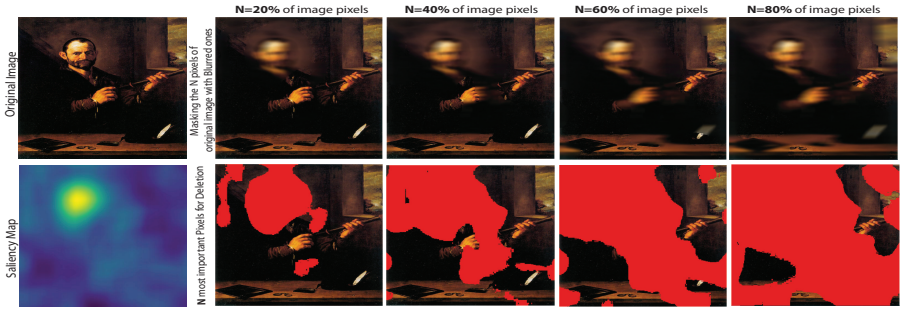


Figure 3: Deletion metric. Examples of N most important pixels (in red) according to the obtained saliency map iteratively removed from the original image and replaced by blurred ones. N could be 30% of image pixels.

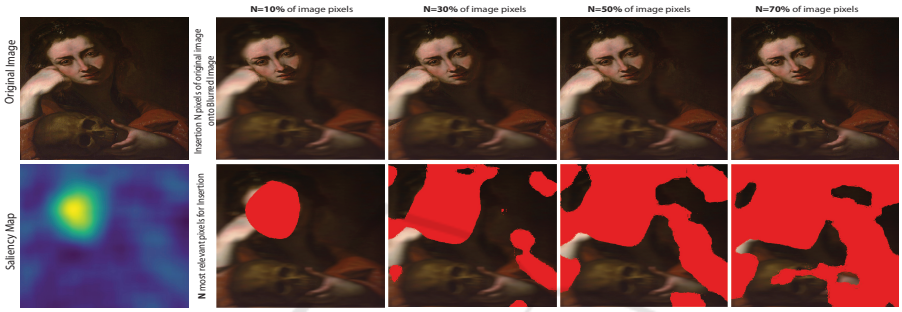


Figure 4: Insertion Metric. Examples of N most relevant pixels (in red) according to the obtained saliency map. N could be 10% of most relevant image pixels, 20%, etc. The N pixels of the original image will be inserted in the blurred image.

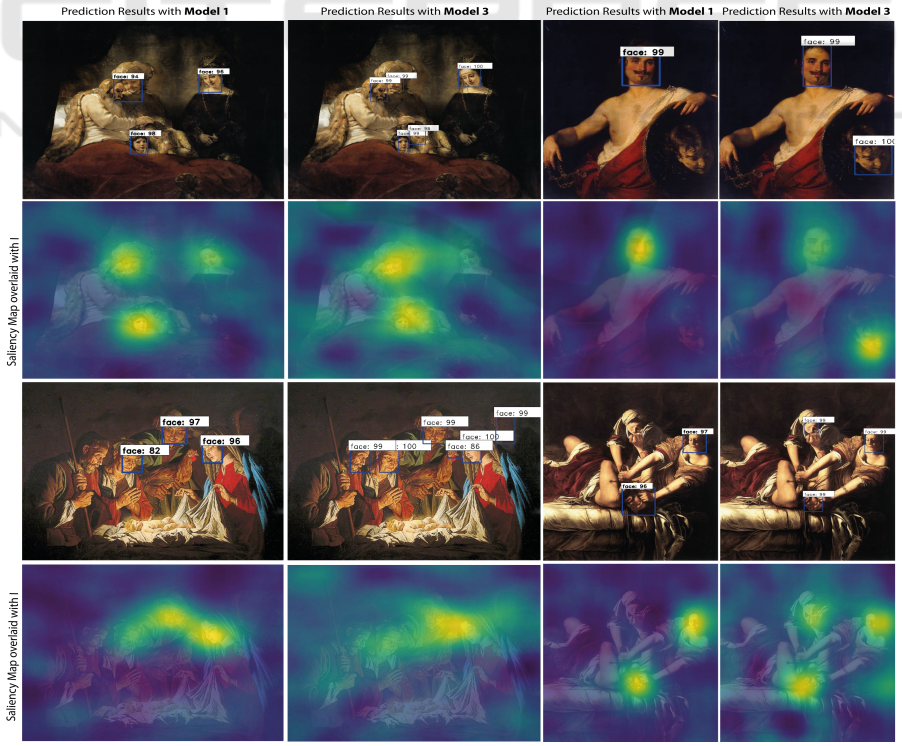


Figure 5: DRISE visualization results for Tenebrism images using Model1 and Model3. The first and third rows are input images with predictions results, others are the input images blended/mixed with their saliency maps.

tion metric iteratively substitutes the N pixels most relevant with original pixels in a blurred version of the original image (see Figure 4). We follow (Zhang et al., 2021) to blur input images by using Gaussian Blur with kernel size = 51 and sigma = 50. We compute then the mean average precision (i.e. accuracy) as the area under precision-recall curve (AUC) to be used for face detection model evaluation.

5.2 Qualitative Results

Figure 5 shows examples comparing the DRISE explanations for bounding boxes predicted by Model1 and Model3. The obtained saliency maps generally seem to highlight the target face with its close surroundings. But we remark that the background surrounding faces in the saliency maps generated for Model1 appears light blue compared to that of the saliency maps generated for Model3. This can indicates that the Model3 explores spatially the image more broadly than the Model1, therefore we can deduce that Model3 has learned more about spatial features from contextual information which could explain performances improvement.

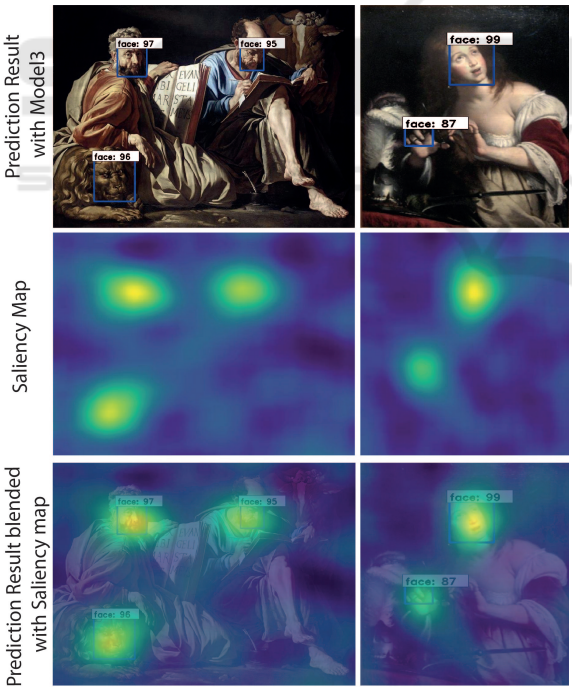


Figure 6: Example of saliency map explanations for false positive detections. A possible interpretation is that important pixels appeared similar to human faces, which could explain the errors.

To understand the reasons of failure detection, we perform model analysis using saliency maps of false positive and false negative detections. Hence, this could provide insights to make improvements to the design architecture of the object detector (e.g. Faster RCNN in our case). Figure 6 shows False positive examples where the Model3 falsely detect faces. The relevant pixels of the saliency map show that the input image contains parts (hands, clothes, animal faces, etc.) which could looks like a face. Hence, we can conclude that the false positive errors start with the feature backbone modeule, and may influence down-stream modules in the face detector architecture namely, region proposal network (RPN) and fast rcnn. Figure 8 shows saliency maps for false negative examples (i.e. missed faces) that Model3 fails to detect them. We remark that the relevant pixels highlighted by the saliency map are considered to be discriminative features for face class, even though the Model3 did not detect them. So the failure is likely occurred while processing these features in future stages in the architecture (potential object region proposal, refining object proposal or duplicate detections suppression).

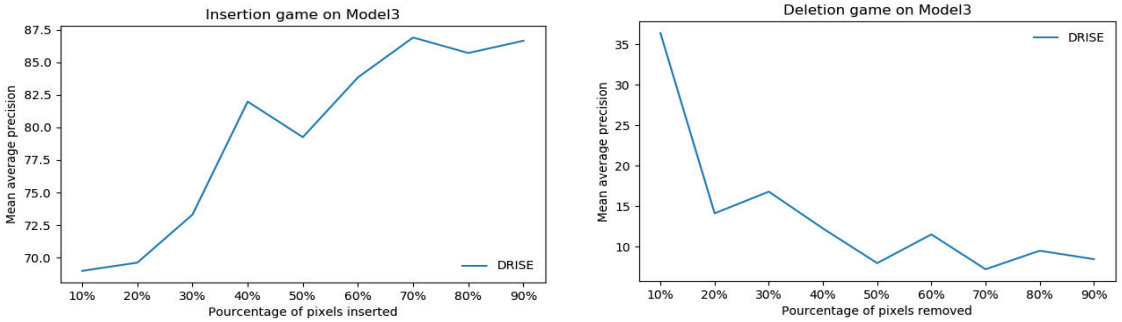
Figure 7: Average face detections (left) and corresponding average saliency maps for Model3 (right).

To provide a more holistic analysis of the common pattern in the model's behaviour across many images, we also compute average saliency map, as described in (Petsiuk et al., 2021), by cropping all face detections from saliency maps, then normalizing them and compute their averages. The obtained results are shown in Figure 7. We remark that this face class has a saliency spread evenly across whole the object (the absence of significant artifact).

5.3 Quantitative Results

We evaluate the performance of DRISE explanation method on Model3 through deletion and insertion experiments. In Figure 9-(a), we show the evolution of the mean average precision w.r.t the percentage of content inserted at each perturbation step. There is an improvement in mean average precision as the 20% of top salient pixels are added back in blurred image. As the insertion of relevant pixels increases, the face detection accuracy improves until reaching 87%. Figure 9-(b) shows the impact of removing relevant pixels on

Figure 8: Explanations for missed detections.



(a) Insertion Test **(b) Deletion Test**

Figure 9: Details of the Insertion & Deletion process.

the model accuracy. As can be seen, when 10% of top most important pixels are removed, there is a sharp drop in the mean average precision to reach around 37%. This continues to drop when we remove pixels in decreasing order of salience (as expected).

6 CONCLUSIONS

In this work, we applied explainability method for object detection analysis and comprehension. An important number of masks is generated to mask input image. Afterwards, the model is running on the obtained masked images to get proposals. Finally, the saliency map is computed as a weighted sum of the masks where weights are pairwise similarities between proposals and target detections. We have successfully demonstrated the application of the DRISE explainability method to our face detector models as a preliminary exploration in Tenebrism painting images. We have also shown its abilities to locate relevant features and discuss reasons for its failures. As further improvement of this work, we should be looking to improve the very processing time consuming of the method.

REFERENCES

Ancona, M., Ceolini, E., Öztireli, C., and Gross, M. (2019). Gradient-based attribution methods. In *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, pages 169–191. Springer.

Brunke, L., Agrawal, P., and George, N. (2020). Evaluating input perturbation methods for interpreting cnns and saliency map comparison. In *ECCV Workshops*.

Cetinic, E., Lipic, T., and Grgic, S. (2019). A deep learning perspective on beauty, sentiment, and remembrance of art. *IEEE Access*, 7:73694–73710.

Chattopadhay, A., Sarkar, A., Howlader, P., and Balasubramanian, V. N. (2018). Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In *2018 IEEE winter conference on applications of computer vision (WACV)*, pages 839–847. IEEE.

Gamra, S. B., Mzoughi, O., Bigand, A., and Zagrouba, E. (2021). New challenges of face detection in paintings based on deep learning. In *VISIGRAPP*.

Hogan, M., Aouf, N., Spencer, P., and Almond, J. (2022). Explainable object detection for uncrewed aerial vehicles using kernelshap. In *2022 IEEE International Conference on Autonomous Robot Systems and Competitions (ICARSC)*, pages 136–141. IEEE.

Ivanovs, M., Kadikis, R., and Ozols, K. (2021). Perturbation-based methods for explaining deep neural networks: A survey. *Pattern Recognition Letters*, 150:228–234.

Kim, V., Cho, H., and Chung, S. (2021). One-step pixel-level perturbation-based saliency detector. In *BMVC virtual conference*.

Mzoughi, O., Bigand, A., and Renaud, C. (2018). Face detection in painting using deep convolutional neural networks. In *ACIVS*.

Nielsen, I. E., Dera, D., Rasool, G., Ramachandran, R. P., and Bouaynaya, N. C. (2022). Robust explainability: A tutorial on gradient-based attribution methods for deep neural networks. *IEEE Signal Processing Magazine*, 39(4):73–84.

Omeiza, D., Speakman, S., Cintas, C., and Weldermariam, K. (2019). Smooth grad-cam++: An enhanced inference level visualization technique for deep convolutional neural network models. *arXiv preprint arXiv:1908.01224*.

Padmanabhan, D. C. (2022). Dext: Detector explanation toolkit for explaining multiple detections using saliency methods.

Petsiuk, V., Das, A., and Saenko, K. (2018). Rise: Randomized input sampling for explanation of black-box models. *arXiv preprint arXiv:1806.07421*.

Petsiuk, V., Jain, R., Manjunatha, V., Morariu, V. I., Mehra, A., Ordonez, V., and Saenko, K. (2021). Black-box explanation of object detectors via saliency maps. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11443–11452.

Pinciroli Vago, N. O., Milani, F., Fraternali, P., and da Silva Torres, R. (2021). Comparing cam algorithms for the identification of salient image features in iconography artwork analysis. *Journal of Imaging*, 7(7):106.

Qiu, L., Yang, Y., Cao, C. C., Liu, J., Zheng, Y., Ngai, H. H. T., Hsiao, J., and Chen, L. (2021). Resisting out-of-distribution data problem in perturbation of xai. *arXiv preprint arXiv:2107.14000*.

Ribeiro, M. T., Singh, S., and 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*, pages 1135–1144.

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. (2017). Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626.

Simonyan, K., Vedaldi, A., and Zisserman, A. (2014). Deep inside convolutional networks: Visualising image classification models and saliency maps. In *In Workshop at International Conference on Learning Representations*. Citeseer.

Smilkov, D., Thorat, N., Kim, B., Viégas, F., and Wattenberg, M. (2017). Smoothgrad: removing noise by adding noise. *arXiv preprint arXiv:1706.03825*.

Springenberg, J. T., Dosovitskiy, A., Brox, T., and Riedmiller, M. (2014). Striving for simplicity: The all convolutional net. *arXiv preprint arXiv:1412.6806*.

Surapaneni, S., Syed, S., and Lee, L. Y. (2020). Exploring themes and bias in art using machine learning image

analysis. In *2020 Systems and Information Engineering Design Symposium (SIEDS)*, pages 1–6. IEEE.

Thakur, S. and Fischmeister, S. (2021). A generalizable saliency map-based interpretation of model outcome. *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 4099–4106.

Tulio Ribeiro, M., Singh, S., and Guestrin, C. (2016). " why should i trust you?": Explaining the predictions of any classifier. *arXiv e-prints*, pages arXiv–1602.

Zeiler, M. D. and Fergus, R. (2014). Visualizing and understanding convolutional networks. In *European conference on computer vision*, pages 818–833. Springer.

Zhang, J., Lin, Z., Brandt, Jonathan, S. X., and Sclaroff, S. (2016). Top-down neural attention by excitation backprop. In *European Conference on Computer Vision(ECCV)*.

Zhang, Q., Rao, L., and Yang, Y. (2021). A novel visual interpretability for deep neural networks by optimizing activation maps with perturbation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 3377–3384.

Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., and Torralba, A. (2016). Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2921–2929.