Black-Box Saliency Map Generation Using Bayesian Optimisation

Mamuku Mokuwe
Centre for Robotics and Future Production
Council for Scientific and Industrial Research
Pretoria, South Africa
mmokuwe@csir.co.za

Michael Burke
School of Informatics
University of Edinburgh
Edinburgh, Scotland
Michael.Burke@ed.ac.uk

Anna Sergeevna Bosman
Department of Computer Science
University of Pretoria
Pretoria, South Africa
anna.bosman@up.ac.za

Abstract—Saliency maps are often used in computer vision to provide intuitive interpretations of what input regions a model has used to produce a specific prediction. A number of approaches to saliency map generation are available, but most require access to model parameters. This work proposes an approach for saliency map generation for black-box models, where no access to model parameters is available, using a Bayesian optimisation sampling method. The approach aims to find the global salient image region responsible for a particular (black-box) model's prediction. This is achieved by a sampling-based approach to model perturbations that seeks to localise salient regions of an image to the black-box model. Results show that the proposed approach to saliency map generation outperforms grid-based perturbation approaches, and performs similarly to gradient-based approaches which require access to model parameters.

Fig. 1. Example of the saliency map (right) and the corresponding input image (left) for class-prediction "papillon". The input image has an overlayed window box, representing an area used to measure sensitivity of the image location to the model's output.

I. INTRODUCTION

Deep learning (DL) techniques have become a standard approach in computer vision. Specifically, the convolutional neural network (CNN) architecture has shown exceptional performance, achieving results comparable to human performance on image recognition tasks [1]–[3]. As a result, CNN models are often deployed in real life as efficient black-box tools. However, there remains a gap in our ability to explain and/or interrogate a model's decisions. Saliency visualisation in machine learning (ML) is a type of visualisation that provides an intuitive explanation of the model's output by highlighting the input regions which contributed the most to the final output. The key motivation behind saliency visualisation in computer vision is to be able to distinguish the image region that contains the information responsible for the model's prediction [4]–[7]. By providing interpretable details about the model's decision, end-user trust can be gained, and introspection can be performed when predictions are incorrect.

Most existing approaches for saliency map generation assume that the model of choice is a white-box, i.e., that certain characteristics of the model are known, and that access to the model's parameters is available [8], [9]. In this work, we introduce a saliency map generation approach that can be used for any black-box model, only requiring access to input and output data, and the ability to query the model. This is particularly important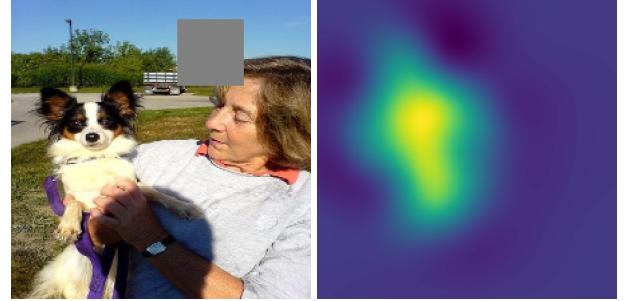 for introspection into third-party tools, or the investigation of models where access to the underlying parameters is not available.

The proposed approach generalises a standard occlusion-based sliding window technique for saliency map generation [6]. The standard occlusion-based approach (referred to as exhaustive search further in this study) involves sequentially blanking regions of the input, and measuring the resultant change in model output. The intuition behind the blanking operation is that for a given model f, an image X, and a partially blanked image $\hat{X}$, the model outputs $f(X)$ and $f(\hat{X})$ should vary significantly if an important feature of X was blanked in $\hat{X}$. An example of a partially blanked image and the corresponding saliency map is shown in Figure 1. Using the exhaustive search approach requires many blanking operations, making it computationally expensive, so typically only very coarse saliency maps can be generated. An important parameter for the exhaustive search is the size of the blanking window. Multiple window sizes can be employed to improve the quality of the saliency map, however, every additional window size further reduces the computational efficiency of the approach.

A method proposed by Burke [10] assumes that the ML model is a black box, and uses probabilistic inference to mimic the behaviour of the model in response to small changes made to the input image. Sampling is performed in areas of interest

to produce the saliency maps. This approach reduces the need to sample at every pixel location, however, there is still a need to manually specify the size of the blanking window used to occlude image regions.

This study builds on the work of Burke [10] to find a general occlusion-based saliency method for black-box models. A Bayesian optimisation sampling approach is used to look for potential regions to blank in an image. Additionally, Bayesian optimisation is used to identify the best blanking window size to use per sample. As a result, the salient parts of the image are sampled more densely, thus reducing the computational cost of producing a saliency map. The primary contributions of this work are as follows:

- We introduce a Bayesian optimisation approach to occlusion-based saliency map generation for black-box models.
- We propose an intersection over union saliency metric to quantitatively measure the accuracy of a saliency map.

The rest of the paper is structured as follows: Section II provides the background on saliency map generation. Section III gives a formal definition of the occlusion-based saliency map generation method. Sections IV discusses the proposed Bayesian optimisation approach to saliency map generation. Section V describes the experimental procedure, and presents the empirical results. Section VI concludes the paper.

II. BACKGROUND

Early work on saliency map generation algorithms [5]–[7] aimed to calculate the gradients from the model to identify pixels that contributed more to the model's output. The derived gradient was accessed using a single pass through the model, and did not take the model architecture, eg. individual network layers, into account. Fong and Vedaldi [11] proposed a method that builds on [5]. Rather than using one pass through the model to obtain the gradient used for the saliency map formation, Fong and Vedaldi's method derives the gradient multiple times using the gradient descent optimisation technique. At each iteration, more information is learned that helps find the size of the perturbation mask which minimises the image classification error. In the reported results, 300 iterations were used to produce the saliency map [11]. In Fong et al. [12], an extension of [11] is presented, which focuses on generating the extremal saliency mask that has a maximal effect on the model's output when applied on the input image compared to other perturbations, i.e., when particular pixels are overlayed on the input image, the model outputs the highest prediction score. These approaches are not applicable to non-differentiable black-box models, and finite difference gradient approximations may be expensive.

More recent approaches [9], [13], [14] have used gradient-based techniques that explicitly manipulate gradients at different network layers. Zhou et al. [13] proposed a saliency mapping approach that requires a direct extraction of the learned weights from the last convolutional layer of a CNN, and applies global average pooling [15] to generate class activation maps (CAM). However, this method can only be used for image classification tasks; thus, a generalisation of CAM was proposed by Selvaraju et al. [14] to extend the approach to accommodate any CNN. Even though these techniques generate saliency maps that are interpretable, they require direct access to the learned features of a CNN, and are designed for the specific CNN-type architecture.

Dabkowski and Gal [16] proposed a method that extracts feature maps at multiple layers and optimises the regions accountable for the output such that (1) the salient region alone allows a sufficient confidence classification, and (2) when the saliency region is removed, the model is not able to predict the correct class of the image. These objectives are further used to define a saliency metric that seeks to identify if the saliency map has indeed discovered the most discriminant region that contributed to a model's prediction.

In Ghariba et al. [17], the encoder-decoder saliency map generation method that requires access to the convolutional blocks of the CNN model was proposed. The technique requires access to five different convolutional blocks to obtain the feature maps and passes the outcome to the five decoder blocks where the features are up-sampled to form saliency maps of the same size as the image. Further, all five attention maps are merged to produce a higher performance map.

Park et al. [18] proposed saliency map generation based on a visual questioning and answering (VQA) model. The saliency maps were used to justify the answer obtained from the model when a question was asked based on the image provided. To generate the saliency maps, access to the model's convolutional blocks and the LSTM features was required.

All of the approaches discussed above require access to model parameters, which makes them unsuitable for black-box model introspection.

A general approach for saliency map generation for black-box models requires testing the model's sensitivity through small alterations to the input, as demonstrated by Zeiler and Fergus [6]. This process sequentially blanks parts of the image region to find input regions that maximally contribute to the final prediction. A variation of this approach was proposed by Burke [10]. The saliency overlay for black-box models proposed in [10] uses a Gaussian process (GP) for saliency map approximation, with a random sampling strategy to search for potential areas to mask in the input image. A limitation of this method is the use of heuristics to select the blanking window size to be used for probing the model. This study proposes a modification of [10] that automatically chooses the appropriate blanking window size using Bayesian optimisation.

The next section provides a formal definition of the exhaustive occlusion-based method for saliency map generation. Subsequently, the proposed Bayesian optimisation approach is discussed in detail.

III. OCCLUSION-BASED SALIENCY

Occlusion-based saliency map generation sequentially blanks patches of input pixels, and evaluates the model's output with respect to the blanked image location. This can be

considered a crude form of sensitivity analysis that identifies salient regions in the image by measuring the change in model output, without tampering with the model.

Without loss of generality, the following definition assumes the model of interest to be a classification model. Let f be the image classification model and X be the images. To evaluate how well f performs given X, we observe the relationship between the blanked image $\hat{X}$ and the change in the model's output, i.e. class probability, $y = f(X) - f(\hat{X})$. The saliency map is formed by the y values, which are obtained by sequentially blanking the image per pixel location, and passing the blanked image $\hat{X}$ to the model.

The exhaustive per-pixel approach requires many model evaluations, where uninteresting regions are also blanked multiple times. This approach is typically infeasible for larger images, and a grid-based approximation is generally proposed for computational feasibility.

The next section presents the proposed Bayesian optimisation approach for saliency map generation.

IV. BAYESIAN OPTIMISATION FOR SALIENCY GENERATION

Bayesian optimisation is a search method that locates the global optima of a black-box function through sequential sampling. The sequential process alternates between two steps: (1) fitting a probabilistic model that approximates the global function value at a point (or region) of f, and (2) using a manipulation function to determine the next best point to evaluate based on the previous mean and variance predictions [19]. This work aims to exploit this method for the purpose of saliency map generation. The primary goal of this work is to generate the saliency map using the search method that chooses globally optimal sampling parameter values for the blanking process, such that the greatest reward is yielded for faster saliency map convergence.

Figure 2 provides a diagrammatic overview of the proposed saliency map generation algorithm using Bayesian optimisation. The black-box model is accessed to extract the information used to generate the saliency map through Bayesian optimisation, which involves fitting a GP model to the observations, and using an acquisition function to select the next region to sample, together with the window size. Section IV-A discusses how the GP model is fit. Section IV-B discusses the acquisition function employed in the study. Section IV-C proposes a region-based ratio metric to be used for the performance evaluation of the saliency maps.

A. Gaussian process

Generating the saliency map for a given image requires iteratively fitting a GP model to determine the posterior probability over functions. A GP by definition is a collection of normally distributed random variables where any finite number of the variables have a jointly Gaussian distribution [20]. Intuitively, the GP is a generalisation of the multivariate Gaussian distribution that is completely specified by the mean function $m(Q)$ and the covariance function $k(Q, Q')$, given a set of observations Q. The GP is used to construct

a probabilistic model g using any finite random collection of data, $\mathcal{D} = \{(\mathbf{x}_{1:i}, y_{1:i})\}$; $\mathbf{x}_{1:i} = \{(u_{1:i}, v_{1:i}, s_{1:i})\}$ and $y_{1:i} = \{f(X)_{1:i} - f(\hat{X})_{1:i}\}$, which is expected to approximate the black-box function f after sufficient observations have been made. Thus, fitting a GP model g requires a finite set of input and output observations from the black-box model f. Let X be an $m \times n$ input image given by a set $\{(u_1, v_1), ..., (u_m, v_n)\}$ of all possible pixel coordinate combinations, and let $S = \{s_1, .., s_l\}$ be the set of variable window sizes of length l. We can fit g using a set of samples $D_{1:i} = \{\mathbf{x}_{1:i}, y_{1:i}\}$. The model g can later be used to predict the approximated saliency map for other unseen observations $Q = \{\mathbf{x}_1, ..., \mathbf{x}_p\}$, $p = m \times n \times l$.

To begin the algorithm, a random point(s) $\mathbf{x}_i \in Q$ is selected, the image X is blanked at coordinates (u_i, v_i) with window variable size s_i, and fed to the model to obtain the output $f(\hat{X}_i)$ before GP evaluation. The finite set $D_{1:i} = \{\mathbf{x}_{1:i}, y_{1:i}\}$, with variables added continuously during the optimisation process, is then used to build the prior GP model g. For GP model parameter choice, the Matérn kernel function is selected. The kernel function is used to estimate the covariance $k(Q, Q')$ of the Gaussian function. Genton [21] discusses different types of kernel functions for ML and their impact. The Matérn covariance function is a kernel function that assumes an uneven feature space. This property introduces flexibility that makes the Matérn function applicable to practical problems [22]. The following equation defines the Matérn function:

$$k_{mat}(\mathbf{x}_i, \mathbf{x}_j) = \frac{\sigma^2 2^{\nu-1}}{\Gamma(\nu)} \left(\frac{\sqrt{2\nu}}{l} r\right)^{\nu} k_v\left(\frac{\sqrt{2\nu}}{l} r\right), \quad (1)$$

where σ is the kernel variance, $r = |\mathbf{x}_i - \mathbf{x}_j|$, $\Gamma(*)$ is the gamma function, $k_v(*)$ is the Bessel function, and $\nu = 2.5$ and $l = 12$ are the hyperparameters of the Matérn function, found using maximum likelihood estimation.
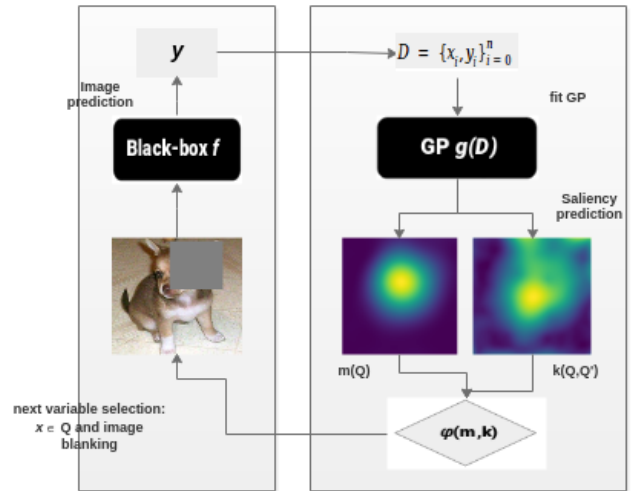

Fig. 2. Bayesian optimisation process for the generation of a saliency map.

To fit the GP iteratively, Bayes' theorem is used, which provides the updated expected solution based on the previously observed variables as follows:

$$p(y|\mathbf{x}) = \frac{p(\mathbf{x}|y)p(y)}{p(\mathbf{x})}, \qquad (2)$$

where $p(\mathbf{x}|y)$ is the likelihood, $p(\mathbf{x})$ is the evidence, and $p(y)$ is the prior. To improve the saliency map, we iteratively update the prior belief of g by adding new observations to D. The new observations are selected such that they improve the model g's knowledge about the true function f or reduce the number of sample variables needed to generate the saliency overlay. To find these observations, we use an acquisition function that uses the posterior mean and covariance values to specify the next pixel coordinates of the image to sample at, together with the window size to blank. Aquisition functions are discussed in the next section.

B. Acquisition functions

In order for the Bayesian optimisation saliency map approach to converge in minimal time, a clever way of selecting regions to blank in the input image that yields the highest reward must be followed. In the Bayesian optimisation approach, an acquisition function φ is used to find the next optimal point $\mathbf{x}_i^* \in Q$ observations, using the predicted posterior mean and covariance [20].

The commonly studied acquisition functions for finding the global optimum of the black-box function can be found in [19]. For this work, we used the expected improvement (EI) acquisition function. The EI acquisition function implicitly balances the exploration-exploitation trade-off, thereby exploring places with high variance, while exploiting at places with low mean [23]. The balance between the exploration and exploitation helps the GP model gain more certainty of the black-box function as a whole.

As shown in Figure 2, the acquisition function is used to compute the next set of elements for the blanking process. The EI acquisition function is given by:

$$\varphi_{EI} = (m(Q)_i - y_i^*)\Phi(Z) - k(Q, Q')_i \phi(Z),$$
$$Z = \frac{m(Q)_i - y_i}{k(Q, Q')_i}, \qquad (3)$$

where $\Phi(Z)$ and $\phi(Z)$ denote the cumulative probability function (CPF) and probability density function (PDF), respectively. The equation uses the CPF and PDF to weigh the strength of either exploiting at the places already visited, or exploring, i.e., trying a new place with the highest uncertainty to see if an improvement on its prediction confidence can be achieved. The value of φ_{EI} is then used to find $\mathbf{x}_i^*$ to evaluate f next, which gives the current optimal y_i^*.

Algorithm 1 summarises the proposed approach.

C. Saliency map evaluation

In order to test the proposed saliency map generation method, and to benchmark it against existing methods, a

---

Algorithm 1 Bayesian optimisation sampling approach for saliency map generation

1: **Start**: Random initialisation of $D_{1:i} = \{\mathbf{x}_{1:i}, y_{1:i}\}$
2: **for** iterations $i = 1, 2, 3, ..., N$ **do**
3:     Fit GP model g using $D_{1:i}$ using Matérn kernel
4:     Predict $m(Q)$ and covariance $k(Q, Q')$ using g
5:     Evaluate $\varphi(Q)$
6:     Select $\mathbf{x}_i^* = \underset{\mathbf{x}}{\arg\min}\, \varphi(Q)$
7:     Query $\mathbf{x}_i$ and then update y_i
8:     Augment dataset $D_{1:i+1} = D_{1:i} \cup (\mathbf{x}_i, y_i)$
9:     Repeat step $[3-8]$ until convergence or the Nth iteration
    return $m(Q)$ and covariance $k(Q, Q')$

---

means of objectively evaluating the quality of the produced saliency maps is necessary. There exist a number of evaluation metrics for saliency map approaches. In Kindermans et al. [24], an approach that assesses the reliability of pre-existing saliency methods was proposed. The method in [24] monitors the invariance of the saliency maps to small changes in the input, and reports the saliency method robustness, with the assumption that minor input changes that do not contribute towards a model's prediction should not alter the saliency map. Dabkowski and Gal [16] proposed a saliency metric that evaluates the quality of the saliency map by finding the smallest rectangle containing the salient region that still allows for correct classification. This metric only considers saliency in terms of the final decision, and would thus penalise a saliency map that included an entire dog over one that simply included the dog's face, since a face is enough to perform classification. In order to address this limitation, we propose a saliency ratio metric r_{sal}, which requires that the bounding box of the desired object be known.

The saliency ratio metric is computed as the sum of the saliency score inside the bounding box over the total sum of the saliency map score. Let $\mathcal{S}$ represent the saliency map for the whole image, and $\mathcal{T}$ the saliency map for the target bounding box. The saliency ratio metric is given by:

$$r_{sal} = \frac{\sum \mathcal{T}}{\sum \mathcal{S}}. \qquad (4)$$

This ratio metric evaluates saliency maps in a manner typically used for segmentation, and rewards saliency within the bounding box specified for a given target, while penalising saliency outside this region. The value of r_{sal} is bounded to the $[0, 1]$ interval, where 0 corresponds to completely missing the bounding box in the saliency prediction, and 1 corresponds to perfectly aligning the saliency prediction with the bounding box.

V. EXPERIMENTAL SETUP AND EMPIRICAL RESULTS

The experiments conducted for this study were designed to test the proposed method on a complex visual task, and to benchmark against the existing saliency map generation approaches. The ImageNet 2012 dataset [25], which mainly
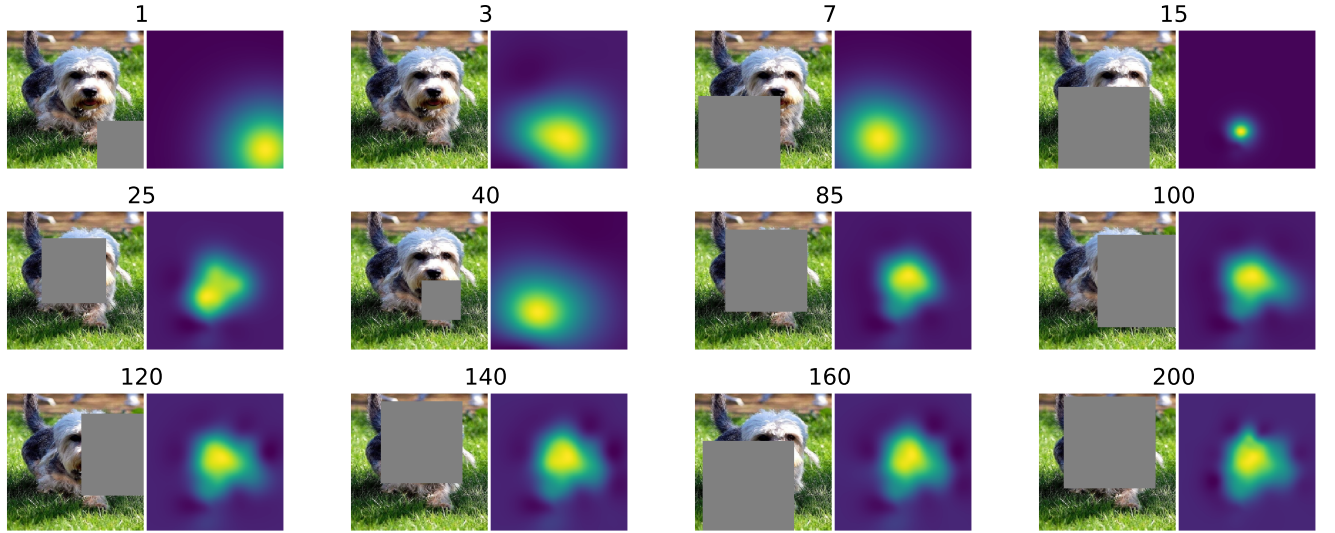
Fig. 3. Illustration of the saliency overlay process using Bayesian optimisation method with a dog as the data sample. The figure shows pairwise input-output results obtained using 12 sample examples. The input images show varied blanking window sizes at different image locations. The output shows the saliency maps, where the blue region indicates pixel regions that are not important, and the green region indicates interesting pixels.

consists of different breeds of dogs, was used to evaluate the performance of the proposed approach. The model used for testing was the pre-trained VGG16 model [2], which is a high-performing CNN architecture. For all images tested, the saliency maps were generated with respect to the associated target class. For the Bayesian optimisation saliency overlay approach, 200 sample steps were used, where at each step a GP model was fit using a set $\mathcal{D} = \{\mathbf{x}_i, y_i\}_{i=1}^{200}$ where $\mathbf{x}_i$ is made of (u_i, v_i) image coordinates, and the blanking window size $s_i \in S, S = \{50, 64, 78, 92, 107, 121, 135, 150\}$. A constant colour 128 for the R, G, B channels was used for the blanked pixels. A total of 500 images were used to quantify the saliency map performance. A convergence analysis showed that 200 samples were typically enough for the optimisation to converge to a maximum on the Imagenet 2012 dataset.

A. Saliency map generation

The results shown in this section illustrate the selection criteria of the Bayesian optimisation approach for the saliency overlay generation. Figure 3 shows a subset of the input-output results, where all the input images are overlayed with a blanking window, and the corresponding output images show the saliency maps at a particular iteration. The blanked position and the blanking window size were specified by the EI acquisition function. The predicted saliency overlays corresponding to the input images shown were produced using the predicted mean of the posterior GP.

Figure 3 shows the automatic selection of the blanking window size, the sampling position, and the saliency maps at specific iteration times. The first saliency map is obtained using a GP model fit with only one sample point, the second with three sample points, and the last with 200 sample points. It can be seen that as the sample size increases, the fitted
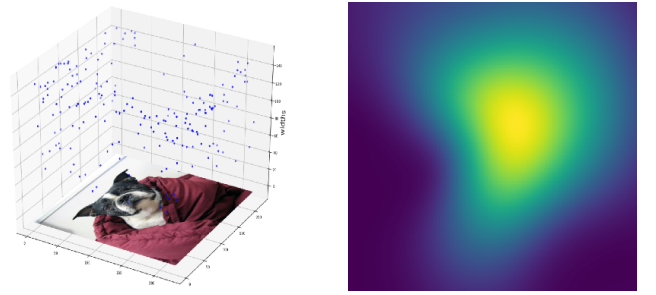

Fig. 4. The saliency map and the points projection used for the saliency map generation.

GP is able to produce a saliency map that can approximate the salient regions of the image using the black-box model, without evaluating the importance at every image pixel. From Figure 3, it can be seen that the saliency map gets better with an increase in sample points, and correctly indicates which image regions contributed more to the output of the given black box model.

B. Exploration-exploitation trade-off

The goal of the proposed saliency map technique is to generate the saliency map from the black-box model with the minimal number of samples, sampling substantially at regions of interest with various (or appropriate) window sizes, as opposed to the exhaustive search saliency map generation approach. Figure 4 shows a 3-dimensional diagram on the left illustrating the sampled points, and the corresponding saliency map to the right. The scattered points represent the (x, y) image coordinates, and the blanking window sizes (z-axis) show the blanking values which were chosen to generate the

Fig. 5. Saliency maps generated from the different methods as follows: (b) the proposed Bayesian optimisation method (Bayes_opt), (c) the exhaustive search discussed in Section III, (d) the saliency by perturbation [11], (e) the grad-CAM [14], and (f) [6], (g) [9] were obtained from [26] toolbox.

saliency map shown on the right of Figure 4. Certain image regions were blanked over more than other regions, where a part of the salient region (i.e. the dog's face) was also sampled substantially. Thus, the proposed technique exploits the area of interest, while still exploring in other regions of the image. With only 200 sample variables, the proposed Bayesian optimisation saliency overlay generation method was able to generate the saliency maps as shown in Figures 3 and 4. The saliency maps produced show that the salient region was correctly identified.

C. Visual quality and interpretability of the saliency maps

The proposed saliency mapping technique is a generalisation of the exhaustive search technique for saliency map generation, thus a comparison of the proposed technique to the exhaustive search technique is necessary. Figure 5 presents a visual comparison of the saliency maps obtained from the proposed Bayesian optimisation approach, the exhaustive search approach, and other saliency map generation methods, applied to the pre-trained image classification black-box model [5], [9], [11], [14]. It is evident from Figure 5 that the saliency maps from the proposed approach and the exhaustive search are comparable, producing saliency maps

with a single prominent attention region. The saliency maps in Figure 5 (d), produced using Fong's method [11], highlight multiple attention regions of very complex shapes, making such salient maps harder to interpret.

By looking at the saliency maps obtained using the proposed method, it can be observed that the dog's head contributed the most as the attention region, since the salient regions are drawn over the region corresponding to the dog's head in the images. Further, the saliency maps produced by the proposed method are better than the saliency maps from the exhaustive search approach, since some of the attention regions on the exhaustive saliency maps (see Figure 5 (b), rows 2, 3, and 6) do not correspond to the salient object of the input image. For example, the image in row two has the object of interest shown in the middle, but the saliency map from the exhaustive search shows the saliency region to be in the bottom-right corner of the saliency map.

Note that the proposed method and the exhaustive search method generate the saliency maps by simply sampling from a black-box model, with no prior information about the learned features. The saliency maps shown in Figure 5 illustrate that this simple technique is sufficient to capture the basic saliency information. The saliency maps produced by the proposed method successfully highlight the interesting/important regions that the model has used to make the class prediction. More complex methods illustrated in Figure 5 (d) - (g) optimise the blanking mask or the gradients that change the classification result of a model. In contrast, the proposed approach and the exhaustive search approach do not require any information about the model parameters or the object of interest from the learned features to produce the saliency map.

D. Localisation of the saliency map

The saliency maps were generated over 500 images, and the saliency ratio metric defined in Section IV-C was computed for each map. The results are shown in Figure 6 for the proposed Bayesian optimisation approach, the exhaustive search, and four other approaches that require access to model parameters. For each bar in Fig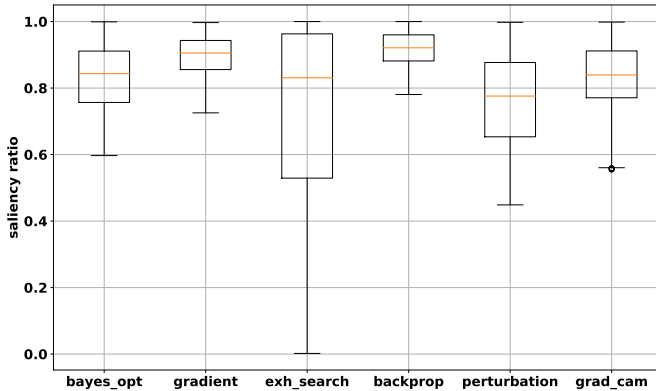ure 6, an overview of how the saliency overlay algorithm performed is shown using the statistical five number summary, namely: the minimum, the first quartile, the mean, the third quartile, and the maximum, together with the outliers represented by circular points below the minimum. In Figure 6, the r_{sal} values of 1 represent perfect saliency maps, and r_{sal} values of 0 correspond to saliency maps that fall outside the target bounding box. The results in Figure 6 show that the proposed Bayesian optimisation saliency method out-performed the exhaustive search and the perturbation method, and was comparable with the gradient-aware methods [5], [9], [11], [14]. The exhaustive search saliency method performed the worst as compared to the other saliency map algorithms. Thus, the Bayesian optimisation approach is a viable technique for saliency map generation, especially when one cannot get access to the model's parameters.

The computational efficiency of the Bayesian optimisation approach is discussed in the next section.

E. Time complexity

The computational complexity of generating the saliency maps using the Bayesian optimisation approach increases cubically with an increase in the number of training variables. The reason behind this is that GP models are computationally expensive, requiring more time as more data is used to fit the model. However, the proposed Bayesian optimisation method generates saliency maps with relatively few samples, making it computationally feasible.

A comparison of the time efficiency of the exhaustive approach and the Bayesian optimisation approach is presented in Table I. Table I shows the time complexity of the saliency map generation, where m, n, l represent the number of the x image values, the y image values and the s blanking window values, respectively, and N is the number of model iterations. The $\mathcal{O}(\cdot)$ notation indicates the number of model evaluations required when performing the experiment. With the exhaustive search approach, time increases linearly with the increase in the number of sample evaluations required to generate the saliency map. For the Bayesian optimisation approach, the time required to run the algorithm increases cubically with an increases in the sample variables. The results presented in Table I show that for larger images, it takes the exhaustive approach more time to generate the saliency map. Therefore, the Bayesian optimisation method requires fewer samples and uses less computational time than the exhaustive search. Thus, the Bayesain optimisation method for saliency map generation is more computationally efficient.

Fig. 6. A comparison of the various saliency mapping methods using the proposed saliency ratio metric, r_{sal}. Same naming abbreviations as in Figure 5 are used for the algorithm names.

TABLE I
COMPUTATIONAL COMPLEXITY

Approach	Time complexity	Num. observations (ImageNet)
Exhaustive	$\mathcal{O}(m \times n \times l)$	401 408
Bayesian opt.	$\mathcal{O}(N^3)$	200

VI. CONCLUSION

In this study, we proposed a method for the generation of saliency maps from black-box models using a Bayesian optimisation sampling approach, generalising an occlusion-based sliding window approach. Instead of sliding a range of blanking windows exhaustively over pixel locations, the proposed approach selects occlusion positions and blanking region sizes based on previous perturbation results. The results obtained show that this approach produces improved saliency maps, and performs similarly to gradient-based approaches, which require access to model parameters. This means that the proposed approach is a useful interpretability aid in black-box model situations, where model introspection is required without parameter access.

In future, additional criteria can be added for blanking. This includes adding variables such as height and width blanking parameters, or different shapes for the blanking window such as triangular or ellipsoid. Selecting the colour to use, as well as choosing the alternative methods to fill the blanking window such as blurring the image, could also be explored. In this work, the GP is used as a proxy model, which requires computing $N \times N$ covariance function manipulations for N observations, making it computationally expensive as N increases. Thus, alternative proxy models such as sparse GP or other approaches could be evaluated as a substitute.

REFERENCES

[1] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1*, ser. NIPS'12. USA: Curran Associates Inc., 2012, pp. 1097–1105. [Online]. Available: http://dl.acm.org/citation.cfm?id=2999134.2999257

[2] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," in *International Conference on Learning Representations*, vol. abs/1409.1556, 2014. [Online]. Available: http://arxiv.org/abs/1409.1556

[3] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–9, 2015.

[4] L. Itti and C. Koch, "A saliency-based search mechanism for overt and covert shifts of visual attention," *Vision Research*, vol. 40, pp. 1489–1506, 2000.

[5] K. Simonyan, A. Vedaldi, and A. Zisserman, "Deep inside convolutional networks: Visualising image classification models and saliency maps," *Computing Research Repository*, vol. abs/1312.6034, 2013. [Online]. Available: http://arxiv.org/abs/1312.6034

[6] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *European Conference on Computer Vision*, 2014.

[7] D. Baehrens, T. Schroeter, S. Harmeling, M. Kawanabe, K. Hansen, and K.-R. Müller, "How to explain individual classification decisions," *Journal of Machine Learning Research*, vol. 11, pp. 1803–1831, 2010.

[8] D. Smilkov, N. Thorat, B. Kim, F. B. Viégas, and M. Wattenberg, "Smoothgrad: removing noise by adding noise," *Computing Research Repository*, vol. abs/1706.03825, 2017. [Online]. Available: http://arxiv.org/abs/1706.03825

[9] J. T. Springenberg, A. Dosovitskiy, T. Brox, and M. A. Riedmiller, "Striving for simplicity: The all convolutional net," *Computing Research Repository*, vol. abs/1412.6806, 2014.

[10] M. Burke, "Leveraging gaussian process approximations for rapid image overlay production," in *Proceedings of the ACM Multimedia 2017 Workshop on South African Academic Participation*, ser. SAWACMMM '17. New York, NY, USA: ACM, 2017, pp. 21–26. [Online]. Available: http://doi.acm.org/10.1145/3132711.3132715

[11] R. Fong and A. Vedaldi, "Interpretable explanations of black boxes by meaningful perturbation," *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 3449–3457, 2017.

[12] R. Fong, M. Patrick, and A. Vedaldi, "Understanding deep networks via extremal perturbations and smooth masks," *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2950–2958, 2019.

[13] B. Zhou, A. Khosla, À. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2921–2929, 2016.

[14] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 618–626.

[15] M. Lin, Q. Chen, and S. Yan, "Network in network," *International Conference on Learning Representations*, vol. arXiv:1312.4400, 2013.

[16] P. Dabkowski and Y. Gal, "Real time image saliency for black box classifiers," *Conference on Neural Information Processing Systems*, 2017. [Online]. Available: https://arxiv.org/abs/1705.07857

[17] B. Ghariba, M. Shehata, and P. McGuire, "Visual saliency prediction based on deep learning," *Information*, vol. 10, p. 257, 08 2019.

[18] D. Park, L. Hendricks, Z. Akata, A. Rohrbach, B. Schiele, T. Darrell, and M. Rohrbach, "Multimodal explanations: Justifying decisions and pointing to the evidence," 06 2018, pp. 8779–8788.

[19] E. Brochu, V. M. Cora, and N. de Freitas, "A tutorial on bayesian optimization of expensive cost functions, with application to active user modeling and hierarchical reinforcement learning," *Technical Report TR-2009-023*, vol. abs/1012.2599, 2010. [Online]. Available: http://arxiv.org/abs/1012.2599

[20] M. A. Osborne, R. Garnett, and S. J. Roberts, "Gaussian processes for global optimization," *Proceedings of the International Conference on Learning and Intelligent Optimization*, 2009.

[21] M. G. Genton, "Classes of kernels for machine learning: A statistics perspective," *The Journal of Machine Learning Research*, vol. 2, pp. 299–312, Mar. 2002. [Online]. Available: http://dl.acm.org/citation.cfm?id=944790.944815

[22] M. L. Stein, *Interpolation of Spatial Data*. Springer, June 1999.

[23] R. Garnett, "Bayesian methods in machine learning - spring 2015: lecture 12: Bayesian optimization," March 2015, washington University in St. Louist - Spring 2015 (Accessed: 10 Mar 2018).

[24] P.-J. Kindermans, S. Hooker, J. Adebayo, M. Alber, K. T. Schtt, S. Dhne, D. Erhan, and B. Kim, "The (un)reliability of saliency methods," *International Conference on Learning Representations*, vol. abs/1711.00867, 2018. [Online]. Available: https://arxiv.org/abs/1711.00867

[25] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet Large Scale Visual Recognition Challenge," *International Journal of Computer Vision (IJCV)*, vol. 115, no. 3, pp. 211–252, 2015.

[26] A. Maximilian, L. Sebastian, S. Philipp, H. Miriam, S. K. T., M. Grgoire, S. Wojciech, M. Klaus-Robert, D. Sven, and K. Pieter-Jan, "iNNvestigate neural networks!" *Computing Research Repository*, vol. abs/1808.04260, 2018. [Online]. Available: https://arxiv.org/abs/1808.04260