SWAG: Superpixels Weighted by Average Gradients for Explanations of CNNs

Thomas Hartley
Cardiff University
hartleytw@cardiff.ac.uk

Kirill Sidorov
Cardiff University
sidorovk@cardiff.ac.uk

Christopher Willis
BAE Systems Applied Intelligence
chris.willis@baesystems.com

David Marshall
Cardiff University
marshallad@cardiff.ac.uk

Abstract

Providing an explanation of the operation of CNNs that is both accurate and interpretable is becoming essential in fields like medical image analysis, surveillance, and autonomous driving. In these areas, it is important to have confidence that the CNN is working as expected and explanations from saliency maps provide an efficient way of doing this. In this paper, we propose a pair of complementary contributions that improve upon the state of the art for region-based explanations in both accuracy and utility. The first is SWAG, a method for generating accurate explanations quickly using superpixels for discriminative regions which is meant to be a more accurate, efficient, and tunable drop in replacement method for Grad-CAM, LIME, or other region-based methods. The second contribution is based on an investigation into how to best generate the superpixels used to represent the features found within the image. Using SWAG, we compare using superpixels created from the image, a combination of the image and backpropagated gradients, and the gradients themselves. To the best of our knowledge, this is the first method proposed to generate explanations using superpixels explicitly created to represent the discriminative features important to the network. To compare we use both ImageNet and challenging fine-grained datasets over a range of metrics. We demonstrate experimentally that our methods provide the best local and global accuracy compared to Grad-CAM, Grad-CAM++, LIME, XRAI, and RISE.

1. Introduction

As Convolution Neural Networks have become more common in sensitive applications such as medical diagnostics [37] or surveillance [38], the more techniques have become necessary to explain a model's predictions. In particular there is a conflict between how well a technique can create explanations that precisely show how a model understands an image, against how well an explanation aligns to interpretable regions within the image. The latter is key to allowing humans to understand an explanation, whilst the former is key to explaining the model accurately. How then does a technique meet these explainability requirements? In particular, is there a technique that can generate explanations which are both accurate and present themselves in an interpretable way?

Doshi-Velez and Kim [8] raised a number of important questions regarding explanations. Important to our work are "cognitive chunks", the basic units of explanations, and task-related factors such as time constraints or whether the explanation is to be local or global. Techniques are often fixed in form and quantity of cognitive chunks present in their explanation, i.e. Class Activation Mapping (CAM) based methods [6, 24, 43] all rely on a coarse feature map taken from the final convolution layer of the network to generate their explanations (i.e. 14×14 or 7×7 for VGG16 and ResNet50 respectively). This results in a large cognitive chunk, which could result in explanations being inaccurate, or at worse, misleading. Gradient based methods such as Deep Taylor [16] or Excitation Backprop [42] are limited in their approach to giving every individual pixel a score with no large-scale spatial coherency. In these approaches, a cognitive chunk is therefore equal to one pixel, too low level a feature to accurately explain a decision in a comprehensible way [39]. Black-box explanation methods such as LIME [21] and RISE [19] are able to vary the number of cognitive chunks (superpixels for LIME, grid cells for RISE) by altering how they generate their perturbation regions. However, this flexibility comes at the cost of having to pass the same image thousands of times through the network. Another limiting factor is that, as the number of cognitive chunks is increased, the number of passes should also be increased accordingly.

To address these issues, we propose a pair of com-

plementary techniques. The first is SWAG (Superpixels Weighted by Average Gradients), a method that allows us to produce explanations using discrete, moderately-sized cognitive chunks, that perform well across a range of datasets and input domains. SWAG uses superpixels as a basis for the cognitive chunks, which are then weighted using the average of the back-propagated guided gradients [30]. However, concerns have been raised about using superpixels as a basis for explanations, as they may not correctly capture discriminative regions [19]. A recent method, XRAI, attempts to alleviate this in two ways. The first is by having multiple sets of overlapping superpixels, the second is by artificially expanding the underlying superpixels by a fixed amount. However, neither of these techniques take into account *how the network is interpreting the image*. We hypothesise that introducing a pixel based saliency map such as backpropagated gradients into the superpixel decision making process will produce regions that better align with the model's use of features. We, therefore, investigate two alternative methods for incorporating saliency maps (based on Simple Linear Iterative Clustering (SLIC) [1]) when creating superpixels. The first method is to modify SLIC so that it not only uses the pixel colour values from the image, but also takes into account the importance of the pixels to the model as determined by a saliency map. The second method is to simply disregard the pixel values from the image and only use the saliency map as a basis for creating the superpixels. The intuition behind generating superpixels using the contributions of a saliency map is that this should cause superpixels to not solely form boundaries between colour regions as with traditional superpixels, but to form boundaries between regions of high and low importance to the network.

This allows us to create superpixels that both describe important image features and the areas important to the network. In this paper, we show that our proposed methods are able to produce explanations that offer improved local and global explanations over other comparable techniques. Our method also has advantages over methods such as LIME, RISE, and XRAI as it only requires a single forward and backward pass to generate an explanation.

2. Related Work

There are multiple ways to create explanations of how a network is behaving. These could broadly be split into observing how the network itself functions, or how the network interprets an input image. Examples of methods that try to explain the the units within a network are Network Dissection [4], Concept Activation Vectors [14] and Activation Maximization [17, 40]. As our proposed method attempts to explain the input space, in this section we focus on similar techniques.

A common method of creating explanations is to back-propagate through the network to the input space. This was first investigated in the works by Zeiler and Fergus [41], and Simonyan *et al.* [27]. Here, the gradients are backpropagated through the network as they would be at training time, except instead of being backpropagated from a loss function, they are backpropagated from the prediction score for the desired class. These techniques were further built upon with the use of guided backpropagation [30], then expanded by combining the gradient with the activations during backpropagation in works such as Layer-wise Relevance Propagation (LRP) [3], Deep Taylor [16], and Excitation Backprop [42]. Integrated Gradients [26] propose that, instead of using a single input, it is better to have a range of scaled inputs (i.e. from zeros to the original input values) and integrate the corresponding gradients.

Using the final activation layer as a basis for explanations has proved to be a popular method. First proposed by Zhou *et al.* [43], Class Activation Maps (CAM) weight and combine the final activation layer using a global average pooling layer. This produces a coarse heat map that centres around the region of the image that is important. This was generalised with Grad-CAM [24] which removed the need for the average pooling layer, instead weighting the activation maps using the mean of the gradient. Grad-CAM++ [6] was later introduced to increase the weak-localisation ability of the method. Finally, creating explanations through the use of multiple perturbations is common. The first example of this for CNNs was the use of a sliding square to occlude regions of the image [41]. As multiple images are passed to the network with regions occluded, a heatmap of how the network output varies due to the occlusion is built. Local Interpretable Model-agnostic Explanations (LIME) [21] uses superpixels as a foundation for the regions to perturb, and then uses the output scores to learn a model to accurately determine the importance of each superpixel to the decision making process. Subsequently, a number of techniques have made use of superpixels as the framework of an explanation. Seo *et al.* [25] and Kapishnikov *et al.* [12] have independently introduced methods of generating explanations produced using multiple levels of superpixels before fusing them together to produce a single explanation. XRAI uses integrated gradients as a method to weight the superpixel regions. As with LIME, this is a computationally expensive method of creating explanations. SHAP, an explanation method based on shaply values, also makes use of superpixels via KernelSHAP [15]. An alternative method for creating explanations using perturbations is Randomized Input Sampling for Explanation of Black-box Models (RISE) [19]. This method generates random masks at a lower resolution than the input and perturbs the input space with these. Whilst the perturbation techniques all have the ability to produce accurate explanations, they are inefficient compared to other methods. This is due to the

multiple passes through the network required, for example LIME typically uses $1{,}000$ passes and RISE uses $4{,}000$ (for VGG16) and $8{,}000$ (for ResNet50).

3. Improved CNN Explanation via SWAG and Gradient-Based Superpixels

Explanations for CNNs can take many forms; however, when explaining how an input image is interpreted by the network, we can broadly split these into either pixel-based (an individual score for every pixel), or region-based (larger regions are used). The popularity of techniques such as Grad-CAM or LIME suggest that region-based techniques offer an increased level of interpretability, that is, they produce an explanation that is easier to understand. However, using these methods comes at the trade-off of the spatial accuracy of the explanation. Pixel-based explanations, where each pixel is individually scored, offer the explanation that can most precisely identify pixels important to the network. Intuitively this makes sense as it has been shown repeatedly through adversarial examples that, by changing only a small selection of pixels, the accuracy of the model can be compromised [32]. Although pixel-based explanations are accurate, they are often described as being of low-quality or less interpretable [35]. In this section, we propose SWAG, a method of weighting superpixels using a backpropagated gradient, and discuss complimentary methods, for generating superpixels using the backpropagated gradient that better capture the discriminative regions of the image.

3.1. SWAG

The core idea of our method is to take gradient values backpropagated to the input and pool them into discrete regions. This requires two separate elements to be generated and combined. To generate the gradients, we use guided-backprop [30] as we found this to perform marginally better when compared to regular backpropagated gradients [27]. This is expanded upon further in a sanity check in Section 4.1.3. For the network architectures of interest this produces an image $M \in \mathbb{R}^{224\times224\times3}$. As per [27], each pixel (i, j) in the final gradient is obtained by taking the max of the absolute values: $M_{i,j} = \max_c |M_{i,j,c}|$, where c is the colour channel. Now each pixel represents a score relating to how important it is to the network's decision.

To define the discrete regions with which to pool the gradients, we use superpixels. We use Simple Linear Iterative Clustering (SLIC) [1], a fast method that accurately adheres to object boundaries [31]. Using SLIC, we are able to control the number of superpixels we generate. Through experimentation we found that starting with 300 superpixels provides a good balance between accuracy and weak-localisation ability (see Section 4.3). With a method for generating superpixels and the values to weight them, we produce an explanation $E_i \in \mathbb{R}$ by weighting each superpixel region R_i (where R_i is the set of pixels belonging to the i^{th} superpixel), with the mean values of M found within that superpixel:

$$E_i = \frac{1}{|R_i|} \sum M \cap R_i. \tag{1}$$

Justification for the use of the mean can be found in the supplemental materials.

3.2. Gradient-Based Superpixels

A number of previous methods have used superpixels as the basis for their explanations [12, 21, 25]. However, it has been noted in previous work [19] that the use of superpixels, whilst aligning well the boundaries of the image, may not align well to regions of the image important to the network. Indeed, in the XRAI method [12], superpixel boundaries are artificially dilated by 5 pixels to incorporate additional edge regions within the superpixel.

In this section, we propose two novel ways of generating the underlying superpixels for use in explanations. This is done by incorporating guided backpropagated gradients as a basis upon which to build the superpixels. We propose creating superpixels using only the backpropagated gradients, as well as a combination of both the image and the gradient.

We first begin by examining how SLIC works using a colour image. SLIC generates superpixels by clustering pixels in both colour and co-ordinate space: $[l_i, a_i, b_i, x_i, y_i]$, where l, a and b represents the CIELAB colour space [7], and x, y are pixel co-ordinates. SLIC proceeds to cluster these to produce cluster centres C_i. Superpixels are allowed to expand or contract within a limited range, in the original SLIC algorithm this is fixed at $2w_s$ from the cluster centre point. Here $w_s = \sqrt{N/K}$, where N is the number of pixels in the image, and K is the desired number of superpixels. To determine whether a pixel (position j) belongs to a given superpixel, its distance to the centre value of the superpixel (at position i) is measured. Here, distance is defined as a combination of both colour distance d_c, and spatial distance d_s:

$$d_c = \sqrt{(l_j - l_i)^2 + (a_j - a_i)^2 + (b_j - b_i)^2}, \tag{2}$$

$$d_s = \sqrt{(x_j - x_i)^2 + (y_j - y_i)^2}. \tag{3}$$

These distances are then combined to give a single distance value D' for each pixel within a superpixel:

$$D' = \sqrt{\left(\frac{d_c}{w_c}\right)^2 + \left(\frac{d_s}{w_s}\right)^2}. \tag{4}$$

Due to the differing scales of d_c and d_s, a scaling component is used for each. For scaling colour distances, a value

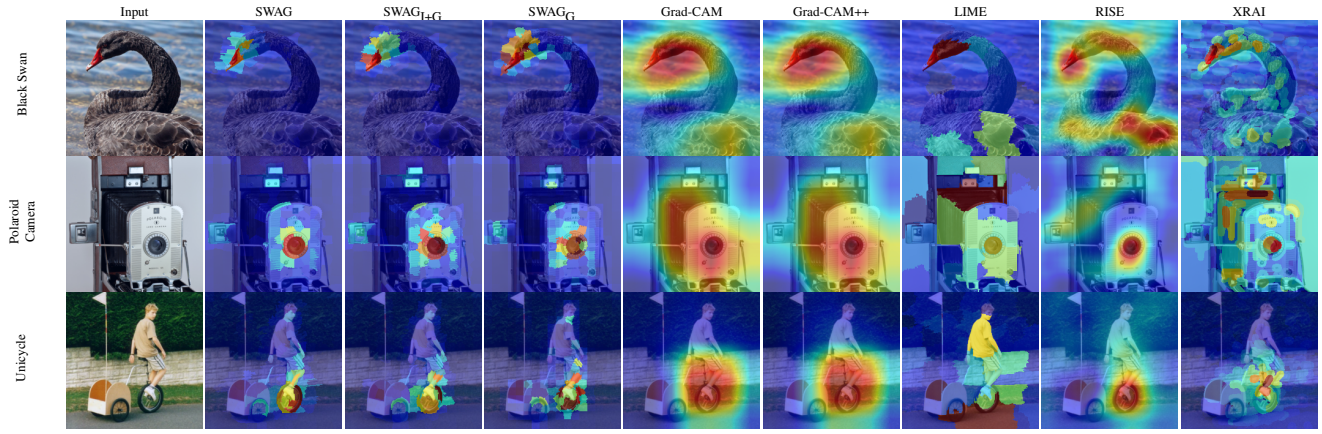

Figure 1: Qualitative comparison between methods using ImageNet and ResNet50. Best viewed in colour. Further examples are in the supplementary materials

w_c is used. When this is large, priority is given to the spatial component, and when it is small, priority is given to the colour distance. The original paper uses a w_c value of 10. Spatial distance is scaled by w_s which seeks to maintain the grid like structure of the superpixels. Clustering proceeds iteratively as in k-means clustering.

Superpixels are designed to adhere to boundaries within an image which makes them useful as a starting point for CNN explainability methods [21]. However, by confining a superpixel method to only taking into account the colour space and distance when generating superpixels, we are potentially creating superpixels in a way that does not lead to producing the most accurate explanations. For example, this process could be splitting an important region of an image across superpixels, when it may be beneficial to have it represented by a single superpixel. We, therefore, propose a method of incorporating a gradient component into the SLIC algorithm. To begin with, we introduce a gradient component g to the initial superpixel description vector: $C_i = [l_i, a_i, b_i, x_i, y_i, g_i]^T$. Here, g is a pixel within our gradient-based explanation M that provides a single score for each pixel. Here, M is scaled between $[0, 100]$ to match the range of LAB values. To compute the distance between pixels and the superpixel centre d_g, as with the spatial and colour distances, we calculate the Euclidean distance: $d_g = \sqrt{(g_j - g_i)^2}$. Following this, we alter the distance function D' to incorporate d_g:

$$D' = \sqrt{\left(\frac{d_c}{w_c}\right)^2 + \left(\frac{d_s}{w_s}\right)^2 + \left(\frac{d_g}{w_g}\right)^2}. \quad (5)$$

We also introduce a new parameter w_g that allows us to control the weighting of the newly introduced gradient element. These give superpixels that are created by combining both the image and gradient, by removing the d_c component we are able to produce superpixels using only the gradient:

$$D' = \sqrt{\left(\frac{d_s}{w_s}\right)^2 + \left(\frac{d_g}{w_g}\right)^2}. \quad (6)$$

Superpixels created using both the image and gradient are labelled with a subscript I+G, whilst those with only the gradient are labelled with a subscript G. For example, SWAG$_{I+G}$ and SWAG$_G$ respectively.

Examples of SWAG using both regular superpixels and our modified methods can be seen in Figure 1 in the columns marked SWAG, SWAG $_{I+G}$, and SWAG$_G$.

4. Image Experiments

In this section, we conduct a number of experiments to examine accuracy (both local and global), weak-localisation ability, and efficiency. We report results across multiple datasets: ImageNet [23], Caltech-UCSD Birds 200(CUB200) [36], Stanford Dogs [13], and Oxford Flowers 102 [18]. Excepting ImageNet, these are all fine-grained datasets, presenting an additional challenge to existing explainability methods where discriminative features may occupy a small region of the image. All work is conducted with PyTorch using pre-trained VGG16 [28] and ResNet50 [10] networks for ImageNet. These models were fine-tuned for the fine-grained datasets for 50 epochs with a learning rate of 0.001 for both VGG16 and ResNet50. Top-1 validation accuracies for VGG16 and ResNet50 respectively are: CUB200 (82.22%, 85.42%), Stanford Dogs (79.60%, 85.09%), and Oxford Flowers (94.95%, 92.24%).

We compare against the following region-based techniques: Grad-CAM, Grad-CAM++, LIME, XRAI, and RISE. We show results using SLIC superpixels generated from both the image and gradient independently, as well as using our combined image and gradient method. Baselines

are often used to evaluate how well a technique performs. In the work by Hooker *et al.* [11], random noise and Sobel edge detection [29] are used as baselines to compare against various saliency map techniques. However, as we are explicitly comparing against region-based explanation approaches, we instead use two additional baselines based on the Euclidean distance from a specific pixel. We use both a centre point Euclidean distance map (referred to as centre), as well as the Euclidean distance to a uniformly randomly chosen pixel (referred to as random). Whilst we believe it is unfair to compare pixel based methods against region based methods due to their inherent precision, we include results for Guided-Backpropagation as it is used for weighting our superpixels. We will see that it provides good local accuracy compared to all region based methods, but poor global accuracy. Our method performs well for both.

4.1. Accuracy (Images)

Explanation accuracy is a measure of how well a method can score regions or pixels of an image important to the network. Whilst some methods have used humans to help evaluate the trustworthiness of an explanation [6, 21], it has been shown that these are vulnerable to confirmation bias [2]. It has also been noted that human-centric evaluations are potentially unreliable as they are measuring how a human interprets the input image, rather than how the network does [19]. To this end, we rely on automatic accuracy measures. Whilst a number of methods for determining accuracy have been proposed, scoring or ranking methods can be inconsistent between different techniques as they can seek to evaluate different aspects of the explanation [33]. In particular, we note the difference between measuring the accuracy for local explanations versus global explanations. Measuring the accuracy of a local explanation allows us to see how well an explanation captures which regions of the input image are important to the network when a specific decision is made. In contrast, global accuracy pertains to how well an explanation is at finding *all* regions of the image that have the potential to influence the network's decision, regardless of whether they are used for the local explanation [8]. We conduct experiments to measure both the local and global accuracy. For a local metric, we use the deletion technique by Petsiuk *et al.* [19]. For the global metric, we use Remove and Retrain (ROAR) from Hooker *et al.* [11].

4.1.1 Local Accuracy: Deletion

In this set of experiments, a saliency value is computed for each pixel. (For techniques that use superpixels, all pixels within a superpixel are assigned the same value.) Pixels are then iteratively removed (by setting to 0). Pixels are deleted in order of importance, most important first. As in Petsiuk's experiment [19], pixels are deleted in batches of 1792 ($\frac{1}{28}$ of the total number of pixels). At each iteration the image is evaluated by the network and the softmax score recorded. We do this for the all the dataset's validation sets and average the softmax scores. Note that to ensure fairness we run all experiments using the original implementations for LIME [22], RISE [20] and XRAI [12]. The only alteration made is to support a PyTorch backend where required. The deletion metric offers a view of what determines local accuracy. It measures both how well a method can determine the regions of the image used by the network and how precise the explanation is. The intuition is that deleting the regions important to a network will force the network to alter its decision. Therefore, as the important regions of the image are deleted, the softmax score will decrease accordingly. This local metric measures the area under the curve (AUC) as features are deleted from the input image. The deletion results can be found in Table 1 and Figure 2.

SWAG is shown to be able to generate better explanations using superpixels generated using either regular SLIC or our modified version across all datasets and models tested, apart from one (Stanford Dogs with ResNet50). We believe that LIME performs well for the Stanford Dogs dataset as the number of superpixels used in LIME seems to better match discriminative regions (typically a dogs head) than SWAG (which tends to focus on the eyes and mouth). Qualitative examples of the Stanford Dogs dataset can be found in the supplemental material. The complementary techniques of SWAG alongside our proposed superpixel methods improve upon all other techniques by a significant margin, except the Dogs / ResNet50 combination.

4.1.2 $\text{SWAG}_{\text{I+G}}$ Optimisation

We use the test for local accuracy as a basis to understand how the colour and gradient weights within $\text{SWAG}_{\text{I+G}}$ influence the performance. We perform a grid search using ResNet50 and Stanford Dogs over the w_c and w_g values from 4 to 20 in steps of 2. We chose these values as in the original SLIC implementation, a value of 10 is chosen. By increasing to 20, it halves the influence of the channel, whilst decreasing to 5 doubles its influence. We found that a w_c value of 20 gave best results. Figure 3 shows how the AUC score alters depending on w_g, reaching a minimum at $w_g = 8$. Further discussion can be found in the supplemental material.

4.1.3 Sanity Check

As a sanity check, we perform the local accuracy measurement using SWAG with the standard image superpixels. In addition to using guided-backpropagation to weight the superpixels, we use random noise, Sobel edges and vanilla gradients. Results in Table 2.

Method	ImageNet		CUB200		Stanford Dogs		Flowers 102	
	VGG16	ResNet50	VGG16	ResNet50	VGG16	ResNet50	VGG16	ResNet50
Random	0.274	0.303	0.296	0.317	0.337	0.371	0.446	0.425
Centre	0.153	0.177	0.153	0.168	0.200	0.233	0.221	0.223
Grad-CAM	0.105	0.142	0.060	0.099	0.097	0.150	0.237	0.235
Grad-CAM ++	0.111	0.147	0.069	0.101	0.104	0.149	0.217	0.234
LIME	0.105	0.125	0.059	0.074	0.087	**0.107**	0.214	0.218
RISE	0.116	0.124	0.057	0.072	0.113	0.129	0.250	0.244
XRAI	0.105	0.137	0.053	0.063	0.090	0.117	0.227	0.188
SWAG	0.092	0.119	0.051	0.062	0.083	0.123	0.206	0.168
$SWAG_{I+G}$	0.084	0.109	0.050	0.060	0.080	0.118	0.195	0.151
$SWAG_{G}$	**0.073**	**0.095**	**0.046**	**0.057**	**0.077**	0.110	**0.177**	**0.137**
Guided-Backprop	0.051	0.074	0.040	0.046	0.042	0.080	0.122	0.086

Table 1: Area under the curve for the deletion experiment. Lower is better. Numbers in bold are the best region based explanations. Note how well the pixel based method performs.

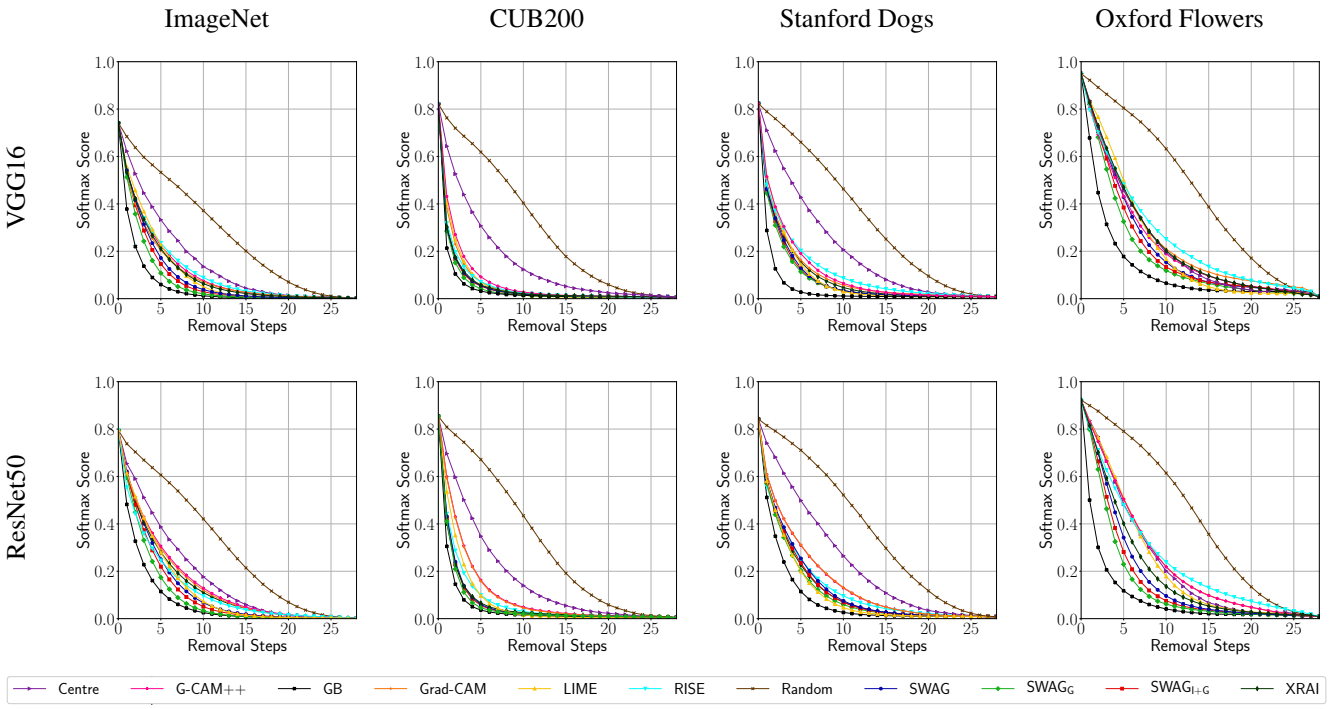

Figure 2: Local accuracy AUC charts. Best viewed in a PDF viewer with zoom ability. Zoomed in graphs featuring the bottom left hand region can be found in the supplemental material.

4.1.4 LIME - Alternative Superpixels

We propose that our method for creating superpixels using the backpropagated gradient can be used as an alternative for other explanation techniques based on superpixels. In this experiment we compute the local accuracy results using LIME with the I, I+G, and G methods for generating superpixels using SLIC with 50 superpixels. Our results using ImageNet are found in Table 3. Note the image only superpixel is different to the scores in Table 1 as here we use SLIC whereas by default LIME uses Quick Shift [34]. From these results we can see that by using our superpixel methods as a drop in replacement for Quick Shift we are able to obtain much better local accuracy results.

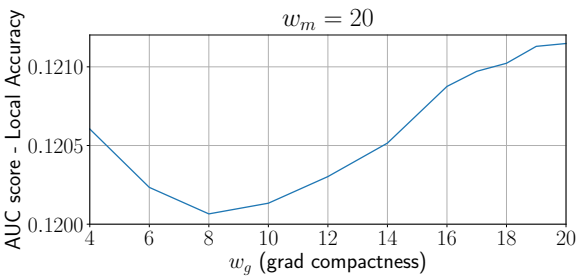

Figure 3: AUC change as we alter the w_g. Lower AUC is better.

Method	Explanation	VGG16	ResNet50
$\text{SWAG}_{\text{Image}}$	GB	**0.092**	**0.119**
	Van	0.134	0.190
	Rand	0.170	0.210
	Sobel	0.154	0.188

Table 2: Sanity check showing how the use of other pixel scoring methods does not perform as well. Lower is better.

Method	VGG16	ResNet50
LIME_{I}	0.107	0.126
$\text{LIME}_{\text{I+G}}$	0.090	0.108
LIME_{G}	0.082	0.099

Table 3: Local accuracy results for LIME. Changing Quick Shift to SLIC and our variants (I+G and G). Lower is better.

4.1.5 Global Accuracy: Remove and Retrain (ROAR)

The previous deletion experiment seeks to gain an understanding of how well a technique explains how a model has learnt to represent a class by the removal of image regions. However, work by Hooker *et al.* [11] suggests that there is a subtlety with this experiment as the images with regions removed are passed back into the network, fall out of the distribution used for training. They argue that it becomes unclear if the performance degradation of a technique in the previous experiment comes from the change in data distribution, or because the technique is genuinely removing important features. They propose an alternative method that requires retraining the network after every stage of feature removal (for removal percentages of [0, 10, 30, 50, 70, 90]). To ensure fairness, they repeat the experiment five times for each method tested. For every explanation technique tested, 30 models are trained, each requiring a new dataset of training and validation images with the correct percentage of pixels removed. Due to the high storage and computational requirements, we are only able to show results for the smaller datasets of CUB200 and Stanford Dogs using ResNet50. We also omit RISE and XRAI as the amount of time taken to generate explanations makes this metric infeasible. We again measure the AUC to obtain a quantitative result, these are shown in Table 4 and Figure 4.

Method	CUB200	Dogs
Centre	0.284	0.393
Grad-CAM	0.219	0.344
G-CAM++	0.218	0.342
LIME	0.210	0.364
SWAG	0.173	0.320
$\text{SWAG}_{\text{I+G}}$	**0.172**	**0.319**
SWAG_{G}	0.179	0.324
Guided Backprop	0.231	0.435

Table 4: ROAR AUC results. Lower is better.

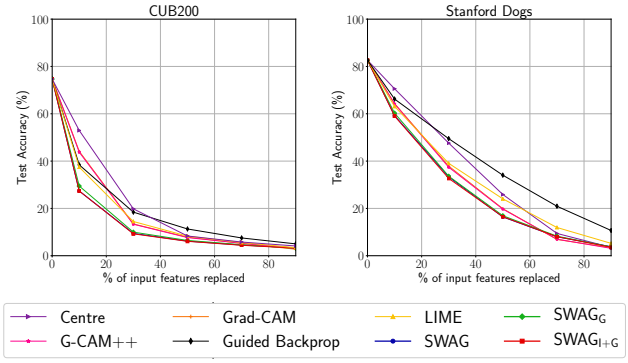

Figure 4: ROAR results. A sharper drop is better.

We see that for both datasets our methods performs better than all the other methods tested at locating features of global importance. Interestingly, despite its strong local accuracy performance SWAG_{G} performs the worst of our proposed methods, with $\text{SWAG}_{\text{I+G}}$ performing the best. There is potential scope for improvement for this score as we tuned it to work better as a local interpretability method through the use of the w_c and w_g values. It is impractical to perform hyper parameter optimisation using the ROAR technique. It is interesting to note that despite its strong local accuracy results, guided-backpropagation performs poorly. This suggests that whilst the gradients can find regions important to the network, they are overly precise to achieve good global accuracy.

4.2. Weak-Localisation (Images)

A common experiment explores an explanations ability to locate a salient object within an image. We used the well-established method [5, 9, 42] of weakly localising the bounding boxes found in the ImageNet validation set. Localisation error is calculated using Intersection over Union

	VGG16			ResNet50		
	Val	Mea	Eng	Val	Mea	Eng
Random	57.43	58.96	57.39	57.43	58.96	57.39
Centre	47.57	48.18	47.68	47.57	48.18	47.68
Grad-CAM	52.06	49.76	51.80	45.94	45.89	44.35
Grad-CAM ++	**47.32**	47.25	46.08	**45.76**	**45.83**	**43.85**
LIME	54.82	52.40	52.82	53.08	50.77	51.19
RISE	55.01	57.94	49.68	52.73	53.82	50.53
SWAG	55.06	46.10	45.01	56.73	52.50	50.65
$\text{SWAG}_{\text{I+G}}$	54.57	46.44	44.95	56.33	52.10	50.70
SWAG_{G}	54.16	**45.95**	**44.86**	56.69	52.07	52.02
Guided Backprop	55.28	46.32	49.63	56.44	51.53	52.35

Table 5: Weak-localisation results as % of localisation error. Lower is better.

(IOU), where an overlap greater than 50% is counted as correct. Implementation details can be found in the supplemental material. While a useful proxy to get insight for the cohesiveness of an explanation, weak localisation experiments do not directly measure the accuracy or quality of an explanation [19]. The results for the weak-localisation experiment are shown in Table 5. For the VGG16 network, we obtain a better overall localisation score than Grad-CAM based methods. Interestingly for VGG16 our method performs better than all others when thresholding using the mean or the energy. However, our method performs poorly when thresholding by pixel value, most likely due to the uneven distribution of values between superpixels. For ResNet50 our method does not perform well when compared to Grad-CAM and Grad-CAM++. Our method mirrors, and sometimes beats guided backpropagation. As it is used to weight, and in some cases define our superpixels, it is possible that using an alternative method could yield better results.

4.3. Effect of Superpixel Count

Varying the number of superpixels alters the performance of SWAG (Figure 5). We observe that increasing the number of superpixels improves the local AUC score. However, whilst the accuracy improves, we note that the ability for SWAG to weakly-localise decreases after 200–300 superpixels. The number of superpixels presents a trade-off between the desired granularity of the explanation and the spatial accuracy (large superpixels can extend beyond the object boundaries, whereas small superpixels cause explanations to become less human-interpretable).

4.4. Efficiency

We measure the mean time to compute an explanation for the first $1{,}000$ images of the ImageNet validation set, cropped to 224×224. Results were computed using an NVidia Titan X GPU. Results are shown in Table 6. From the results we see that there is a wide gap between techniques such as LIME, RISE and XRAI compared to other

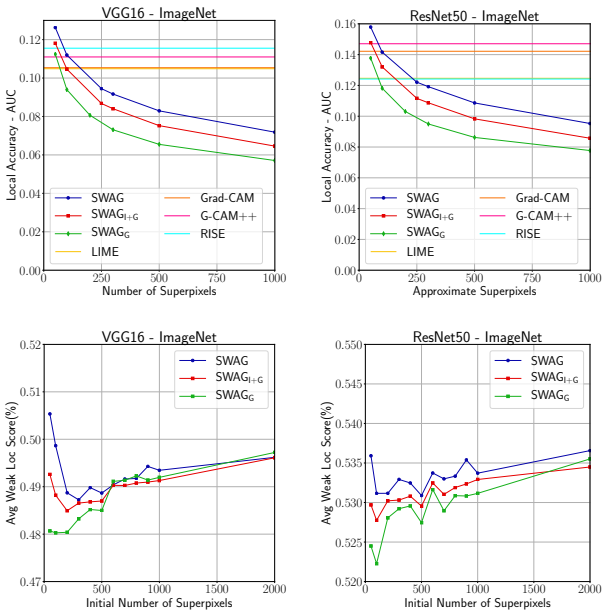

Figure 5: Variation in local accuracy (top row) and weak localisation (bottom row) over a range of superpixels.

Method	VGG16	ResNet50
Grad-CAM	0.03	0.03
Grad-CAM++	0.03	0.03
LIME	5.80	4.76
RISE	13.19	17.48
XRAI	31.10	30.57
SWAG	0.12	0.18
$\text{SWAG}_{\text{I+G}}$	0.12	0.18
SWAG_{G}	0.07	0.10

Table 6: Mean computation time in seconds

gradient based methods. Whilst our technique is marginally slower than Grad-CAM or Grad-CAM++ we note that a much *higher accuracy* is achieved *for only a minimal loss of efficiency*.

5. Conclusion

In this paper we have introduced a complementary pair of novel techniques for weighting superpixels with guided gradients and for generating superpixels that better match discriminative regions within an image. We have shown the technique to be effective for both local and global explanations on a range of image datasets.

6. Acknowledgements

This work is generously funded by BAE Systems and the EPSRC via iCASE award 1852482.

References

[1] Radhakrishna Achanta, Appu Shaji, Kevin Smith, Aurelien Lucchi, Pascal Fua, and Sabine Süsstrunk. SLIC superpixels compared to state-of-the-art superpixel methods. *IEEE transactions on Pattern Analysis and Machine Intelligence*, 34(11):2274–2282, 2012.

[2] Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity checks for saliency maps. In *Advances in Neural Information Processing Systems (NIPS)*, pages 9525–9536, 2018.

[3] Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLoS ONE*, 10(7), 07 2015.

[4] David Bau, Bolei Zhou, Aditya Khosla, Aude Oliva, and Antonio Torralba. Network dissection: Quantifying interpretability of deep visual representations. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3319–3327, July 2017.

[5] Chunshui Cao, Xianming Liu, Yi Yang, Yinan Yu, Jiang Wang, Zilei Wang, Yongzhen Huang, Liang Wang, Chang Huang, Wei Xu, Deva Ramanan, and Thomas S. Huang. Look and think twice: Capturing top-down visual attention with feedback convolutional neural networks. In *The IEEE International Conference on Computer Vision (ICCV)*, pages 2956–2964, December 2015.

[6] Aditya Chattopadhyay, Anirban Sarkar, Prantik Howlader, and Vineeth N. Balasubramanian. Grad-CAM++: Generalized gradient-based visual explanations for deep convolutional networks. In *IEEE Winter Conference on Applications of Computer Vision*, pages 839–847, 2018.

[7] Commission International de L'Eclairage. Colorimetry. *CIE Pub*, 15(2):29–30, 1986.

[8] Finale Doshi-Velez and Been Kim. Considerations for evaluation and generalization in interpretable machine learning. In *Notions and Concepts on Explainability and Interpretability*, pages 3–17. Springer, 2018.

[9] Ruth C Fong and Andrea Vedaldi. Interpretable explanations of black boxes by meaningful perturbation. In *The IEEE International Conference on Computer Vision (ICCV)*, pages 3449–3457, Oct 2017.

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, June 2016.

[11] Sara Hooker, Dumitru Erhan, Pieter-Jan Kindermans, and Been Kim. A benchmark for interpretability methods in deep neural networks. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 9734–9745. Curran Associates, Inc., 2019.

[12] Andrei Kapishnikov, Tolga Bolukbasi, Fernanda Viegas, and Michael Terry. Xrai: Better attributions through regions. In *The IEEE International Conference on Computer Vision (ICCV)*, October 2019.

[13] Aditya Khosla, Nityananda Jayadevaprakash, Bangpeng Yao, and Li Fei-Fei. Novel dataset for fine-grained image categorization. In *First Workshop on Fine-Grained Visual Categorization, IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Colorado Springs, CO, June 2011.

[14] Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, and Rory sayres. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV). In *International Conference on Machine Learning (ICML)*, volume 80 of *Proceedings of Machine Learning Research*, pages 2668–2677. PMLR, 10–15 Jul 2018.

[15] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems (NIPS) 30*, pages 4765–4774. Curran Associates, Inc., 2017.

[16] Grégoire Montavon, Sebastian Lapuschkin, Alexander Binder, Wojciech Samek, and Klaus-Robert Müller. Explaining nonlinear classification decisions with deep Taylor decomposition. *Pattern Recognition*, 65:211–222, 2017.

[17] Anh Nguyen, Alexey Dosovitskiy, Jason Yosinski, Thomas Brox, and Jeff Clune. Synthesizing the preferred inputs for neurons in neural networks via deep generator networks. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 3387–3395. Curran Associates, Inc., 2016.

[18] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *Proceedings of the 2008 Sixth Indian Conference on Computer Vision, Graphics and Image Processing*, pages 722–729, 2008.

[19] Vitali Petsiuk, Abir Das, and Kate Saenko. RISE: randomized input sampling for explanation of black-box models. In *British Machine Vision Conference 2018, BMVC*, 2018.

[20] Vitali Petsiuk, Abir Das, and Kate Saenko. RISE: randomized input sampling for explanation of black-box models. `https://github.com/eclique/RISE`, 2018.

[21] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "Why should I trust you?": Explaining the predictions of any classifier. In *22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1135–1144, 2016.

[22] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "Why should I trust you?": Explaining the predictions of any classifier. `https://github.com/marcotcr/lime`, 2016.

[23] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015.

[24] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-CAM: Visual explanations from deep networks via gradient-based localization. In *The IEEE International Conference on Computer Vision (ICCV)*, pages 618–626, Oct

2017.

[25] D. Seo, K. Oh, and I. Oh. Regional multi-scale approach for visually pleasing explanations of deep neural networks. *IEEE Access*, 8:8572–8582, 2020.

[26] Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. Learning important features through propagating activation differences. In *34th International Conference on Machine Learning (ICML)*, volume 70 of *Proceedings of Machine Learning Research*, pages 3145–3153. PMLR, 06–11 Aug 2017.

[27] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. In *2nd International Conference on Learning Representations, ICLR 2014, Workshop Track Proceedings*, 2014.

[28] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations (ICLR)*, 2015.

[29] Irwin Sobel and G. Feldman. A 33 isotropic gradient operator for image processing. *Pattern Classification and Scene Analysis*, pages 271–272, 01 1973.

[30] Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin A. Riedmiller. Striving for simplicity: The all convolutional net. In *3rd International Conference on Learning Representations, ICLR 2015, Workshop Track Proceedings*, 2015.

[31] David Stutz, Alexander Hermans, and Bastian Leibe. Superpixels: An evaluation of the state-of-the-art. *Computer Vision and Image Understanding*, 166:1–27, 2018.

[32] Jiawei Su, Danilo Vasconcellos Vargas, and Kouichi Sakurai. One pixel attack for fooling deep neural networks. *IEEE Transactions on Evolutionary Computation*, 23(5):828–841, 2019.

[33] Richard Tomsett, Dan Harborne, Supriyo Chakraborty, Prudhvi Gurram, and Alun Preece. Sanity checks for saliency metrics. In *Thirty-Fourth AAAI Conference on Artificial Intelligence*, 2020.

[34] Andrea Vedaldi and Stefano Soatto. Quick shift and kernel methods for mode seeking. In *European conference on computer vision*, pages 705–718. Springer, 2008.

[35] Jorg Wagner, Jan Mathias Kohler, Tobias Gindele, Leon Hetzel, Jakob Thaddaus Wiedemer, and Sven Behnke. Interpretable and fine-grained visual explanations for convolutional neural networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9089–9099, June 2019.

[36] P. Welinder, S. Branson, T. Mita, C. Wah, F. Schroff, S. Belongie, and P. Perona. Caltech-UCSD Birds 200. Technical Report CNS-TR-2010-001, California Institute of Technology, 2010.

[37] Nan Wu, Krzysztof J. Geras, Yiqiu Shen, Jingyi Su, S. Gene Kim, Eric Kim, Stacey Wolfson, Linda Moy, and Kyunghyun Cho. Breast density classification with deep convolutional neural networks. In *ICASSP*, pages 6682–6686, 2018.

[38] Dan Xu, Yan Yan, Elisa Ricci, and Nicu Sebe. Detecting anomalous events in videos by learning deep representations of appearance and motion. *Computer Vision and Image Understanding*, 156:117–127, 2017.

[39] Mengjiao Yang and Been Kim. Benchmarking Attribution Methods with Relative Feature Importance. *CoRR*, abs/1907.09701, 2019.

[40] Jason Yosinski, Jeff Clune, Anh Nguyen, Thomas Fuchs, and Hod Lipson. Understanding neural networks through deep visualization. *arXiv preprint arXiv:1506.06579*, 2015.

[41] Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *European Conference on Computer Vision*, pages 818–833. Springer, 2014.

[42] Jianming Zhang, Sarah Adel Bargal, Zhe Lin, Jonathan Brandt, Xiaohui Shen, and Stan Sclaroff. Top-down neural attention by excitation backprop. *International Journal of Computer Vision*, 126(10):1084–1102, 2018.

[43] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2921–2929, June 2016.