Classifier-Based Evaluation of Image Feature Importance

Paper id: 23

xxx

Abstract

Significant advances in the performance of deep neural networks have created a drive for understanding how they work. Different techniques have been proposed to determine which features (e.g., CNN pixels) are most important for the classification. However, these techniques have only been judged subjectively by a human. We address the need for an objective measure to assess the quality of different feature importance measures. In particular, we propose measuring the ratio of the CNN's accuracy on the whole image compared to an image containing only the important features. We also consider scaling this ratio by the relative size of the important region in order to measure the conciseness. We demonstrate that our measures correlate well with prior subjective comparisons of important features, but importantly do not require their usability studies. We also demonstrate that the features that multiple techniques agree are important have higher impact on accuracy than those features that only one technique finds.

Contents

1 Introduction

There has been tremendous advancement in the performance of deep neural networks (DNNs), specifically in the task of image recognition using deep convolution networks (CNNs) Krizhevsky *et al.* [2012]. However, due to the complexity of the models, there is much interest in understanding and explaining how the networks work. A variety of visualization techniques have been proposed to indicate which pixel features are most discriminative for CNNs to determine their classification prediction on a given image (e.g., Selvaraju *et al.* [2016]; Simonyan *et al.* [2014]; Zintgraf *et al.* [2016]). For example, Simonyan *et al.* uses the gradients to determine which pixels determine classification. Zintgraf *et al.* proposes occluding the image systematically and observing the confidence scores to determine the discriminative pixels.

With such different algorithms to determine pixel "importance", we are interested in comparing the regions that each one finds. Most current techniques to evaluate visualizations are qualitative Zhang *et al.* [2016]; Zintgraf *et al.* [2016] and use human studies Selvaraju *et al.* [2016] to determine which regions people believe are most discriminative. However, people's opinions of important features may be different than what the CNN actually uses to determine its classification. Additionally, the small number of people used in the studies leads to challenges in replicability. In contrast to subjective measures, one recent objective method was proposed to evaluate important pixels by perturbing a random subset and then reevaluating the perturbed images to observe how the CNN prediction confidence changes Samek *et al.* [2016]. However, the non-determinism in the perturbations could introduce new artifacts in the image which might confuse the classifier rather than evaluate pixel importance.

In this work, we also propose that feature importance should be measured objectively with respect to the predicting CNN. Like Samek *et al.*, our goal is to evaluate whether different techniques identify the pixels within the image that most affect the accuracy of the CNN. However, unlike the prior work that measured the decrease in accuracy caused by adding noise to the important region, we contribute metrics that begin with an uninformative baseline image and measures the increase in accuracy caused by adding the important pixels. We call this metric Simple Confidence Gain (SCG). Our second metric - Concise Confidence Gain (CCG) - builds upon SCG by measuring conciseness of the region of pixels required to classify the image correctly.

Using our metrics, we contribute comparisons of three different algorithms for finding important regions of images on two different datasets - Place365 Zhou *et al.* [2016] and our own dataset containing images of various floors in our building. The results from our metrics are internally consistent and correlate well with prior subjective comparisons of important features. However, we note that this may not always be the case because people may not know what pixels are important.

Finally, as we evaluated different algorithms, we noticed that many of the important pixels identified by those algorithms differ. We contribute a technique to find more concise important pixel regions by identifying the pixel regions that are in agreement between different algorithms (i.e., the region intersections). These intersecting regions trade-off CNN accuracy with conciseness and represent an alternative approach to reducing the important region size. We conclude that our metrics can be used in conjunction with, or even to replace, the qualitative evaluation using human studies to evaluate new importance regions and visualization techniques.

2 Related Work

A variety of deep network visualization techniques have been developed to understand CNNs (e.g., Zintgraf *et al.*; Selvaraju *et al.*; Lengerich *et al.*). We roughly divide these techniques into two categories - class model visualization and image specific visualization. Class model visualizations such as Simonyan *et al.*; Yosinski *et al.* aim to understand how the neurons in the network contribute to the classification. Image specific visualization techniques aim to find what features (pixels) the CNNs find most informative Zeiler and Fergus [2014]; Selvaraju *et al.* [2016]; Zhang *et al.* [2016]; Simonyan *et al.* [2014]; Zintgraf *et al.* [2016].

In this work, we focus on evaluating image specific visualizations for the important features that they highlight. We refer to the feature-finding algorithms as *importance functions*. For example, Simonyan *et al.*; Zhang *et al.* have developed error backpropagation-based techniques to find the importance of different regions of an image for a prediction by computing gradients with respect to the image. The work has been extended to evaluate activations of particular neurons rather than pixels in images. Zeiler and Fergus have developed a technique for sensitivity analysis by occluding patches in the image. Zintgraf *et al.* [2016] has also created a procedure for finding importance function using occluding patches. While we evaluate our new metrics on the three importance functions, our metrics can be applied to any techniques that find features that are important to classifiers.

There are relatively few evaluation techniques for evaluating or comparing importance functions. Selvaraju *et al.* use human studies to compare different importance function's ability to discriminate between classes. However, human studies only evaluate the quality of the function's visualization from a human's point of view and do not give any measure of how well the function has captured what the network has learned. Samek *et al.* propose an algorithm to objectively evaluate importance functions by randomly perturbing a small region around the important pixels of the image and observing the confidence scores from the classifier. They do this random perturbation sequentially in order of importance for 100 relevant pixels. The confidence scores during the process are then used for comparing different importance functions. However, since the work randomly perturbs the pixels of the images, it could introduce new artifacts in the image which might confuse the classifier. We propose new metrics for objectively evaluating importance functions. Our proposed metrics are 100 times faster than the prior approach and are deterministic.

3 Importance Functions

We first formalize the definition of an importance function before proposing metrics for evaluating them. We assume that a CNN classifier C outputs $p(I = y|w)$, the probability of an image $I \in [0,1]^{c*N}$ with c channels (i.e., 3 for RGB) and N pixels having classification y given the trained weights w. For clarity, we will refer to the ith pixel in the image as $I[i]$. Given C and I, an importance function importance(I, C), takes as input I and C, and outputs a *heat map* $H \in [0,1]^N$ that contains a measure of relevance of each pixel $I[i]$ to the class y. A variety of importance functions, each with their own heat map, have been proposed for explaining the classification predictions of CNNs. In this section, we briefly summarize three existing visualization techniques that we will use for evaluating our new metrics.

Occluding Patches (*occ*), Simonyan *et al.* [2014] The idea behind the approach is if a key feature in an image gets occluded, then the classifier's confidence will fall. The heat map rates the regions that cause large confidence drops as more important than surrounding pixels. Specifically, a gray square patch of a fixed size, called the occlusion mask, is used to systematically occlude parts of the input image. The *occ* algorithm first creates a visibility mask V as the inverse of the occluded patch: $V[i] = 0$ if $I[i]$ is occluded and 1 otherwise.

By weighing the non-occluded regions of the image by their classifier accuracy, the algorithm combines all visibility mask scores to generate the heat map H. The high confidence regions are those pixels that have high values in heat map H.

$$H = 1 - \frac{\sum_{j=1}^{J} p(I_{o,j} = y|w) * V_j}{J} \tag{1}$$

where $I_{o,j}$ is the j^{th} occluded image and J is the number of images generated by systematic occlusion. Note that the heat map H is a function of the size of the occluding patch, so we can evaluate different sized patches to understand how their resulting heat maps change our importance measures.

Gradients (*grad*), Zeiler and Fergus [2014] For the gradient visualization technique, H represents the magnitude m of the derivative of the classification confidence with respect to the image. The magnitude of ith pixel m_i represents the sensitivity of the network's prediction to the change in that pixel's value and is equal to the derivative of the classification probability $p(I = y|w)$ with respect to $I[i]$. We expect the classifier accuracy to be more sensitive to the change in values of the important features than others. Note that since the gradients are pixel-wise importance values for the image, the heat map is generally of high entropy thus lacks continuous important image regions.

Contrastive Marginal Winning Probability (*C-MWP*), [Zhang *et al.*, 2016] For a CNN acting on I, C-MWP models H using probabilistic Winner-Take-All(WTA) formulation Tsotsos *et al.* [1995]. The WTA identifies the neurons that are relevant to the task in a particular layer using *Excitation Backpropagation* that computes the Marginal Winning Probabilities (MWP). After identifying the relevant neurons, the heat map is generated using the most relevant neurons' receptive field– pixels the neuron acts on. The MWP heat map's discriminative ability can be improved by backpropagating contrastive signals to produce C-MWP heat maps. Contrasting signals for an image belonging to class A, is the difference in the gradients of A and *not* A classifier.

4 Analyzing Important Features

Given an image and classifier that determines what the image contains, our goal is to understand which pixels of the image are most important to its classification. Because different algorithms may determine different pixels as important, we are interested in a measure of goodness to compare different important regions. Prior work has focused on allowing users to rate visualizations overlayed on the image. In contrast, we propose to reduce variability in subjective preferences by contributing measures that utilize the classifier itself. This proposal also captures the relevance of the important regions to the classifier which is not captured in the human studies.

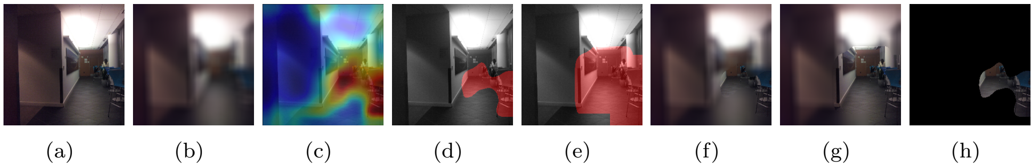


Figure 1: (a) an original image, (b) base image obtained from using Gaussian kernel G_k, (c) heat map obtained using *C-MWP* where red and blue represents the most and the least important pixels, (d) binary mask obtained after thresholding the heat map for top ρ=5% pixels, (e) mask obtained after growing the regions of the mask in (d). (f) and (g) are the hybrid images created using mask in (d) and (e) respectively using base image (b). (h) is the hybrid image obtained using mask in (d) and a base image obtained using zeros kernel Z_k.

4.1 Problem Formulation

Rather than visualizing heat maps for use by humans to judge whether the importance function has found important parts of the image (e.g., Heat map Figure 1(c) for Figure 1(a)), we use those heat maps to identify a subset of pixels that, when added to a baseline image, closely matches the accuracy of the CNN classification compared to the original image.

We compute a binary mask M as shown in Figure 1(d), that signifies whether each pixel is included in the important region or not. A mask $M \in \{0,1\}^N$ is created such that each pixel i takes value: $M[i] = 1$ if important and 0 otherwise. In this work, we threshold the top ρ highest value pixels of the heat map as our mask but other techniques such as graph cut Boykov and Jolly [2001] could also be used.

Our goal is to add those pixels to a base image to capture the accuracy of the important region compared to the original image. We define a base image I_K which represents the image I altered using a kernel function K. The kernel is chosen such that it renders the pixels comparatively less informative for classification. In this work, we have explored two such kernels: a Gaussian kernel G_k, which blurs the image refer Figure 1(b), and a zeros kernel Z_k, with zeros in its all element values. A *hybrid image* $I_{M,K}$ as shown in Figures 1(f) and 1(h) contains important pixels from the mask M and less informative base image pixels using kernel K otherwise.

Next, we propose two measures of confidence gain to reflect the proportion of confidence that can be attributed to the important region compared to the original image. Our metrics yield high values when the important regions are responsible for a majority of the confidence in the original image.

4.2 Metric: Simple Confidence Gain (SCG)

The Simple Confidence Gain (SCG) measures the ratio of the improvement in accuracy of the base image to the hybrid image containing only important features compared to the improvement in accuracy of the base image to the original image. Note that we assume that the kernel is predefined and the same for all compared masks M.

$$SCG(I, K, M) = \frac{p(I_{M,K} = y|w) - p(I_K = y|w)}{p(I = y|w) - p(I_K = y|w)} \tag{2}$$

We calibrate $p(I = y|w)$ with respect to $p(I_K = y|w)$ to measure only the relative increase in the classification accuracy due to the important features and not the transformed non-important regions. SCG outputs values from 0 to 1. A value of SCG close to 1 indicates that the masked pixels contribute highly to the classifier accuracy. Values close to 0 indicate that the mask has little contribution.

4.3 Metric: Concise Confidence Gain (CCG)

The Concise Confidence Gain (CCG) builds upon SCG in two ways. First, it requires the important region to produce an accurate classification. Second, it measures the conciseness of the important region necessary to classify the image correctly. The idea of CCG is to increase the region under M to form a new accurate mask AM as shown in Figure 1(e), such that the classifier predicts the class y of the hybrid image $I_{AM,K}$ correctly as shown in Figure 1(g). There are many ways of increasing the size of the mask. For example, we can simply construct a new mask from the heat map with an increased threshold ρ. In this work, we have chosen to grow the mask using the dilate operation which enlarges the boundary regions of the foreground pixels. With the new hybrid image, the CCG metric is calculated as:

$$CCG(I, K, AM) = \frac{\big(p(I_{AM,K} = y|w) - p(I_K = y|w)\big) * N}{\big(p(I = y|w) - p(I_K = y|w)\big) * A_{AM}} \tag{3}$$

where A_{AM} is the area of the image masked by AM. Note that two different masks that are originally the same size need not be the same after dilation, as it depends on the geometry of

Algorithm 1 Procedure for calculating CCG

```
Input: $H$, $I$, $K$, $\rho$, $w$, $y$
$M \leftarrow$ GetBitMask(H, $\rho$) /* Creating the mask */
$I_K \leftarrow$ TransformImage(I, K) /* Creating the base image */
/* Loop until prediction matches the correct class */
repeat
  $I_{M,K} \leftarrow$ CreateHybrid(M, I)
  y' = Classify($I_{M,k}$, w) /* Predict for the hybrid image */
  /* Break if the predicted the correct class */
  if y' == y then
    $I_{AM,k} \leftarrow I_{M,k}$
    $A_{AM} \leftarrow$ TotalElements(M) /* Area of the mask */
    break
  end if
  M $\leftarrow$ DilateGrow(M) /* Grow the mask */
until
CCG $\leftarrow$ Compute with Equation 3
```

the mask. We divide the relative confidence by the ratio A_{AM} to image size N. The complete algorithm for finding CCG is shown in Algorithm 1.

Unlike SCG, CCG values can range from 0 to N. High values of CCG reflect both 1) high accuracy of the hybrid image $I_{AM,K}$ compared to the original image, and 2) conciseness of the mask AM compared to the size of the whole image. Unlike SCG, it can be used to compare features of different sizes. In a sense, CCG measures the density of information in a region that can sufficiently determine the class, while SCG measures the total information in a feature set.

4.4 Agreement Between Importance Functions

In practice, many importance functions find very dissimilar important regions, which raises the question whether regions that are in agreement between algorithms are more informative than those that are not. In other words, pixels that several different importance functions can agree are important are potentially more likely to be the most discriminative and have higher values of our CCG metric compared to the regions in individual importance functions. We do not use SCG metric for the comparisons as the size of the mask changes.

There are many different ways to find the intersection of important regions. For example, it is possible to add two heat maps and then segment it to generate a new mask. In this work, we take the intersection of the binary-masked regions. We will compare the accuracy of the individual importance functions to the features that the functions have in common.

Applications Beyond Visual Domains. Our metrics make it possible to compare different algorithms more directly without user studies. The metrics could be used to analyze measures other than classification accuracy by substituting its value - $p(I = y|w)$ - with a term that captures the new measure. Additionally, unlike the visualization techniques, our metrics apply not only to visual domains but also to non-visual domains as long as an importance function exists. Finally, we note that SCG and CCG could be used to evaluate which pixels are important in an incorrect classification, i.e., by analyzing the important regions in the image with respect to the incorrect label rather than the ground truth label.

		$\rho = 25\%$							
		SGC*100				CCG*100			
		Floor		Places365		Floor		Places365	
Config		G_k	Z_k	G_k	Z_k	G_k	Z_k	G_k	Z_k
occ	10	43	34	31	23	114	98	93	77
	50	28	38	22	21	103	96	85	82
	100	36	27	18	18	105	94	81	80
grad	0	46	24	39	19	113	70	110	82
	2	61	31	43	20	107	74	112	88
	5	57	30	44	25	116	88	116	90
C-MWP		**71**	**39**	**50**	**37**	**120**	**113**	**137**	**115**
grad+occ		25	30	20	15	223	139	122	89
C-MWP+grad		43	32	28	16	**239**	169	**155**	**122**
C-MWP+occ		29	30	17	13	225	**171**	154	114
		$\rho = 5\%$							
occ	10	25	22	20	16	161	132	103	88
	50	16	19	14	13	135	136	96	86
	100	20	22	11	12	119	113	91	84
grad	0	18	22	26	14	163	70	128	83
	2	35	**28**	**29**	15	**237**	98	128	94
	5	31	22	28	18	189	122	130	96
C-MWP		**40**	22	27	**21**	208	**162**	**162**	**125**
grad+occ		6	18	7	9	209	166	133	103
C-MWP+grad		16	22	11	9	**330**	230	**186**	**137**
C-MWP+occ		7	15	4	5	249	**269**	179	133

Table 1: Average SCG and CCG (*100) for individual masks with $\rho = 25\%$ and 5%. C-MWP performs best (bold) in almost all datasets, base image kernels, and ρ values. The intersection of all pairs of importance functions were also tested. Bold values show that CCG is higher for the pairs than the best single important region. For the pairs of functions, patch size for *occ* is 10 and dilation for *grad* is 5.

5 Experiments

We performed experiments on two different datasets and the three importance functions. We first describe our datasets and the corresponding CNNs. Then, we present our experiments for evaluating the different important functions using our proposed measures. Finally, we use our metrics and the datasets to compare the accuracy of the individual importance masks to those features that different masks have in agreement.

5.1 Datasets and Classifiers

We chose a scene recognition task because it is challenging for a person to determine which part of an image is most important to classification compared to object detection tasks in which the object within the image should be most important. We expect subjective analysis of scene recognition visualizations to be less consistent because there are many areas of the image that may affect scene classification.

Building-Floor. When a robot navigates across many floors of a building, it can be challenging to determine which floor it is currently on. We collected the Building-Floor dataset in one of our buildings. Each image contains the scene just outside the elevator from six different floors of the building. The goal of the classifier trained on this dataset is to recognize the floor the image belongs to.

For each of the floors in the building, ten images were taken at specific locations that our robot stops at outside the elevators, with slight variation in position. To simplify the analysis, all the images were taken at the same time of the day, and the effects of people moving around in the building are not considered. The training data consists of three images, and remaining seven images form the testing dataset.

In order to classify the floor for each image, we chose to use a CNN based on Siamese architecture Bromley *et al.* [1993], because it has been shown to perform well in one-shot learning problems Koch *et al.* [2015]. Our training network of nine layers followed AlexNet Krizhevsky *et al.* [2012] with an input size of N=227x227, in a modified Siamese architecture proposed in Sun *et al.*; Zheng *et al.*, which combined the identification– Softmax, and the verification loss– Contrastive, for better performance. We combined identification and verification loss with a pre-trained network to reduce overfitting which could happen when the complexity of network is higher than that of the data. During training, the first seven layers of our network were initialized from Places205-AlexNet which was trained in the Places205-Standard dataset and provided by the authors Zhou *et al.* [2014]. The remaining two layers were trained from scratch. During training, the contrastive loss was utilized in the eighth layer which is a dense layer of 1000 units, while the softmax loss was employed in the ninth layer. During testing, our network was able to classify all the images in the dataset correctly.

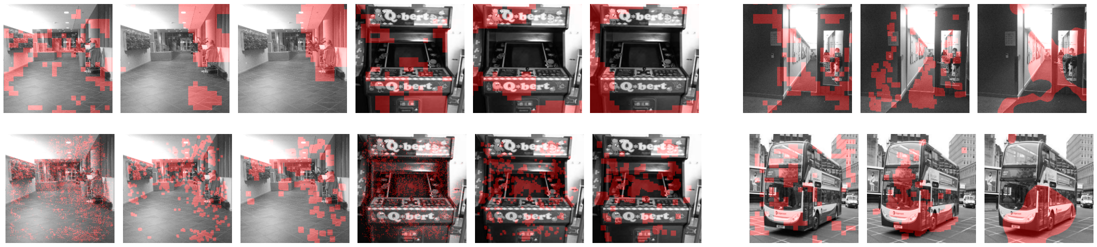
Places365-Standard. The Places365-Standard dataset Zhou *et al.* [2016] contains indoor and outdoor images from 365 categories. We used the Places365-AlexNet model with an input size of N=227x227, provided by the authors, which was trained using ∼1.8 million images. For testing, 200 random images were selected from the dataset without any other consideration like ground truth label.

5.2 Experimental Procedure

The three importance functions each required chosen parameters. For *occ*, the heat map is a function of the size of the occluding patches. For our evaluation, we varied the size of the occluding patches $\in \{10, 50, 100\}$ pixels. *grad*'s heat map H is of high entropy, so we dilate the raw heat map 0, 2, and 5 times with a 3x3 kernel. Dilating smoothens the heat map and improves the continuity of important regions as shown in Figure 2(a). For *C-MWP*, we use H from the *pool*2 layer for both of the networks as we lose the spatial accuracy at higher layers. For our *C-MWP* implementation, we used the source code provided by the authors.

Given the heat maps generated by each importance function, we then created the binary masks using simple thresholding to ensure that ρ% (ρ = 5% and 25%) of the top features are consistent across tests. To create the base images, we used two different techniques - a Gaussian kernel G_k of size 17*17 for creating the blurred base images and a zero kernel Z_k to substitute black for the non-important pixels. In order to create the accurate hybrid image, we grow the regions of the mask using a 3x3 dilate operation. When testing the common features in agreement between importance functions, we evaluated all pairs on both datasets, namely *grad+occ*, *C-MWP+grad*, and *C-MWP+occ*. For the experiments in agreement with *occ*, we have fixed the patch size to be 10, and for *grad*, the number of dilating operations is 5.

During the experiments, the images where $p(I = y|w) - p(I_K = y|w)$ and $p(I_{M,k} = y|w) - p(I_k = y|w)$ are less than zero are not considered, as it violates our assumption that K renders



(a) (b)

Figure 2: (a) The image masks for the (top) *occ* importance function and (bottom) *grad* importance function generated with the parameters ρ=25% and with dilation = {0, 2, 5} respectively for one image from the Building-Floor dataset (left three images) and one image from the Places365 dataset belonging to the class *amusement station* (right three images). (b) A side by side comparison of occ(patch size = 10), *grad*(dilation = 5), and *C-MWP* respectively on an image from the Building-Floor and the Place365 dataset (ρ=25%).

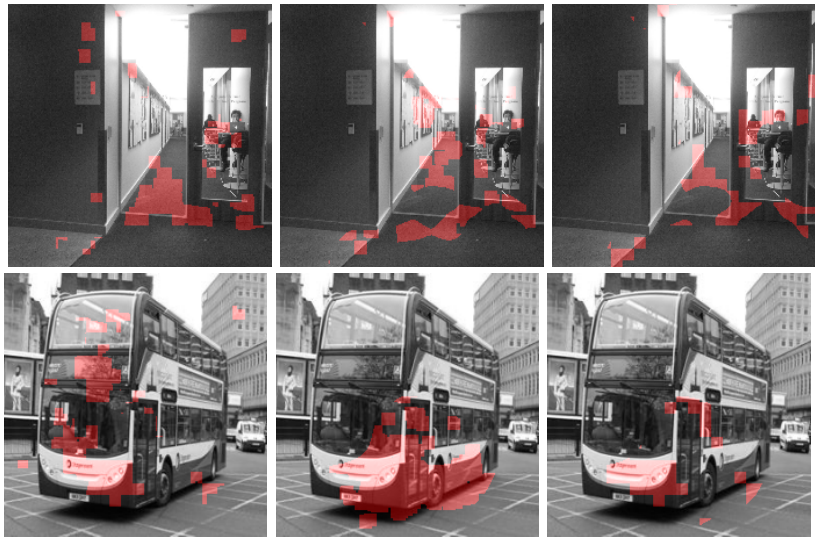


Figure 3: A side by side comparison of the three pairs of importance functions (*grad+occ*, *C-MWP+grad*, and *C-MWP+occ* respectively) on an image from the Building-Floor dataset and the Place365 dataset (ρ=25%).

the pixels less informative for the classifier. An example for such a case could be an image belonging to the class *night*, using a zero kernel Z_k will make it the best image for the class. In both, the datasets less than 5% of the images violates this assumption. Additionally, some images were omitted for a specific test condition. During CCG calculation, we ignore the images where the M had to be grown more than 50 times in order to avoid growing the mask too much beyond the original mask. In practice, this was an issue for some test images with ρ=5%, about 5% of the test images were omitted from each dataset with this condition.

6 Results And Discussion

Evaluating Importance Functions. In total, 38 and 180 images were used for testing in Building-Floor and Place-365 dataset respectively. The quantitative results of the evaluation of the individual importance function's masks are shown in Table 1. The masks obtained from varying the respective parameters for *occ* and *grad* on one image from each dataset are shown in Figure 2(a)(top) and 2(a)(bottom). For *occ*, we find that the patch size of 10 on average performs better than 50 which is better than 100. *occ* rates all pixels covered by large occlusion patches as important when only a small area under the occlusion may actually be important. A patch size of 10 occludes smaller regions of important features and are more concise, and hence better capture what the network has learned. For *grad*, dilating the region 2 and 5 times performs as well or better than not dilating the heat map. When the important regions have high entropy (e.g., 0 dilations), the hybrid image is not informative for the classifier.

A side by side comparison of the mask obtained for *occ* with patch size=10, *grad* with number of dilations=5, and *C-MWP* are shown in Figure 2(b). Among the three, on average *C-MWP* performs the best, followed by *grad* according to both metrics, which can also be qualitatively seen in Figure 2(b) with a bus station example. In the figure, *C-MWP* captures the more discriminative features like the bus wheel and the floor for the image belonging to the bus station while the other techniques capture non-relevant regions like the buildings. Our metrics consistently find that the C-MWP importance function outperforms the others across a random sample of scenes which demonstrate the robustness of our metrics to large variations in features and image labels.

Effect Of Varying Parameters. We varied $\rho = 5\%$ and 25% as the size of the mask. We find that *SCG* metric was higher for larger masks and *CCG* was higher for smaller ones. This indicates that when the percentage of the image decreases so does the size of important regions leaving a more concise area. However, the larger mask size does not increase the mask proportionately. CCG finds that the smaller area is sufficient for achieving "high enough" accuracy while SCG values the higher accuracy achieved with more pixels. Selecting values of ρ for future evaluations of importance functions should take this tradeoff into account. CCG on average dilates the accurate mask for 10 ($\rho = 5\%$) and 6 iterations (25%) respectively. Although we did not analyze kernel parameters extensively, our initial experiments showed that the parameter did not significantly affect the metrics nor relative ranking between the importance functions.

Agreement Between Importance Functions. We then analyzed the features that were in agreement or in common between different importance functions (results shown in Table 1 and the masks are visualized in Figure 3). The in-common features resulted in higher average CCG values than those computed using individual important features, as predicted. However, the in-common average SCG values are lower than the individual ones, because SCG considers only the amount of information gained and not the density or conciseness of the features like CCG.

There are far fewer discriminative features that pairs of importance functions have in common (Figure 3) compared to those found by individual functions (Figure 2(b)). This result is most apparent in the Building-Floor dataset, where the in-common importance masks have captured the glass door and the hallway behind it while the individual masks have captured other features as well. Subjectively, we agree that the glass door is the most discriminative feature of that image and of the floor because it is the only floor with a glass door. Similarly, in the image from the Places365 dataset, in-common importance masks capture only the bus while the individual functions capture additional features.

Among all combinations of importance functions for agreement, *C-MWP+grad* performs better than *C-MWP+occ* by a small margin. And, both *C-MWP+grad* and *C-MWP+occ* outperform *grad+occ* by a large margin. Because *C-MWP* performs the best individually followed by *grad*, it makes sense that the features in agreement would also perform better. We have also conducted experiments to find the intersection of all three importance masks and observed that the CCG scores were on an average higher and SCG scores were lower than the ones of common importance masks for two importance functions.

We conclude that finding common features from different importance functions result in a more concise region of important features, which can be beneficial for preventing information overload to humans for visualization as well as determining the most discriminative pixels for classification.

Quantitative vs Qualitative Evaluation. Finally, we compared our metrics to prior findings based on subjective results[1]. Our metrics find that *C-MWP* outperforms the other two by a significant margin, followed by *grad* and then *occ*. This result *matches* the conclusion in Zhang *et al.* which compares *C-MWP* and *grad* qualitatively and shows that the former is able to produce better localization of objects.

However, it may not necessarily always be the case that our objective metrics match the subjective results, when the classifier uses pixels that are different than what a human would find important. We emphasize that our metrics capture the features that the DNN actually uses to classify, and therefore the direct correlation or comparison to subjective approaches is not necessarily possible.

7 Conclusion

Prior work to evaluate regions of images that are most discriminative for classification has been largely subjective, depending on humans to rate visualizations. In this work, we contribute two metrics – SCG and CCG – to address the need for an objective measure to assess the quality of different feature importance measures. The principle behind both the metrics is to compare the proportion of the classifier accuracy that is attributed to the important features. Our CCG metric also takes into account the conciseness of the region and requires the classifier to classify the image accurately.

We have used the metrics to compare three visualization techniques on two scene recognition datasets and demonstrate the differences between the metrics and importance functions with different parameters. Our results correlate with prior subjective evaluations, although this result is not guaranteed. We demonstrated that the features that appear in multiple importance functions result in higher CCG scores, i.e., they are a more concise set of better discriminative features than a single importance function. We conclude that our metrics can be used in conjunction with or in lieu of subjective human studies to objectively evaluate importance functions towards understanding and explaining CNN classification.

References

Yuri Y Boykov and M-P Jolly. Interactive graph cuts for optimal boundary & region segmentation of objects in nd images. In *Computer Vision, 2001. ICCV 2001. Proceedings. Eighth IEEE International Conference on*, volume 1, pages 105–112. IEEE, 2001.

[1] The prior results were completed on object recognition tasks.

J. Bromley, I. Guyon, Y. LeCun, E. Säckinger, and R. Shah. Signature verification using a siamese time delay neural network. In *Proceedings of the 6th International Conference on Neural Information Processing Systems*, pages 737–744. Morgan Kaufmann Publishers Inc., 1993.

G. Koch, R. Zemel, and R. Salakhutdinov. Siamese neural networks for one-shot image recognition. In *ICML Deep Learning Workshop*, 2015.

A. Krizhevsky, I. Sutskever, and G. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

B. Lengerich, S. Konam, E. Xing, S. Rosenthal, and M. Veloso. Visual explanations for convolutional neural networks via input resampling. In *Workshop on Visualization for Deep Learning, Thirty-fourth International Conference on Machine Learning*, 2017.

W. Samek, A. Binder, G. Montavon, S. Lapuschkin, and K. Müller. Evaluating the visualization of what a deep neural network has learned. *IEEE Transactions on Neural Networks and Learning Systems*, 2016.

R. Selvaraju, A. Das, R. Vedantam, M. Cogswell, D. Parikh, and D. Batra. Grad-cam: Why did you say that? visual explanations from deep networks via gradient-based localization. *CoRR*, abs/1610.02391, 2016.

K. Simonyan, A. Vedaldi, and A. Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. In *ICLR Workshop*, 2014.

Y. Sun, Y. Chen, X. Wang, and X. Tang. Deep learning face representation by joint identification-verification. In *Advances in neural information processing systems*, pages 1988–1996, 2014.

J.K. Tsotsos, S.M. Culhane, W.Y.K. Wai, Y. Lai, N. Davis, and F. Nuflo. Modeling visual attention via selective tuning. *Artificial intelligence*, 78(1-2):507–545, 1995.

J. Yosinski, J. Clune, T. Fuchs, and H. Lipson. Understanding neural networks through deep visualization. In *In ICML Workshop on Deep Learning.* Citeseer, 2015.

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

J. Zhang, Z. Lin, J. Brandt, X. Shen, and S. Sclaroff. Top-down neural attention by excitation backprop. In *European Conference on Computer Vision*, pages 543–559. Springer, 2016.

Z. Zheng, L. Zheng, and Y. Yang. A discriminatively learned CNN embedding for person re-identification. *CoRR*, abs/1611.05666, 2016.

B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva. Learning deep features for scene recognition using places database. In *Advances in neural information processing systems*, pages 487–495, 2014.

B. Zhou, A. Khosla, A. Lapedriza, A. Torralba, and A. Oliva. Places: An image database for deep scene understanding. *CoRR*, abs/1610.02055, 2016.

L. Zintgraf, T. Cohen, and M. Welling. A new method to visualize deep neural networks. *CoRR*, abs/1603.02518, 2016.