Analysis and Evaluation of Grad-CAM Explanations

Rajmund Nagy
rajmundn@kth.se

Doumitrou Daniil Nimara
nimara@kth.se

Livia Qian
liviaq@kth.se

Abstract

In this project, we reimplement the paper *Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization* from 2016 which introduced a visual explanation method for convolutional neural networks. Our experiments focus on evaluating the reproducibility of the results shown in the paper (e.g. localization task, pointing game, comparison with occlusion maps); moreover, we propose novel experiments in order to better understand the strengths and weaknesses of this technique. In this regard, we 1) analyze Grad-CAM's ability to explain chest X-Rays (medicine is a field in which localization is of utmost importance) and compare its localization capability with other explanation methods; 2) measure its *fidelity* and *contrastivity*; and 3) introduce a new metric (to the best of our knowledge) based on the notion of *sensitivity*. Our results advocate for Grad-CAM's efficacy in CNNs and provide new information regarding its particularities; for instance, we show that great network performance does not translate as smoothly to good localization in the more specialized medical dataset (where we achieve results comparable with other papers). Furthermore, our implementation of Grad-CAM++ provides a promising alternative, outperforming Grad-CAM in the aforementioned difficult dataset. Lastly, our fidelity experiments propose that the method might get outperformed by non-CNN based explanation methods when a large portion of the network is non-convolutional.

1 Introduction

Despite the increase in the commercial use of deep learning, many neural networks are still treated as black boxes. This is particularly problematic in tasks where mistakes are exceptionally costly (e.g. self-driving cars). Many visual explanation methods have been developed in recent years to tackle this issue. In this project, we investigate the paper *"Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization"* [1] from a reproducibility perspective and carry out three new experiments to further evaluate the proposed technique. Our code is publicly available here [1].

The remainder of this report is structured the following way: Section 2 will introduce related work in order to better conceptualize Grad-CAM's reasoning, approaches and strengths. We will then summarize Grad-CAM in more detail in Section 3. After providing a firmer understanding of the method, we will analyze the original paper's reproducibility in Section 4. In Section 5, we will explore new experiments to examine Grad-CAM in novel ways, both quantitatively and qualitatively. Finally, we will summarize our findings and share our greatest challenges in Sections 6 and 7.

2 Related work

In 2016, Zhou et al. [2] showed that global average pooling layers (GAP) can help CNNs retain their ability to localize objects despite being only trained for image classification. They proposed Class Activation Maps (CAM), which visualize a network's attention on a given image when making

[1] GitHub Repository

HT20 Deep Learning Advanced, DD2412, KTH Royal Institute of Technology.

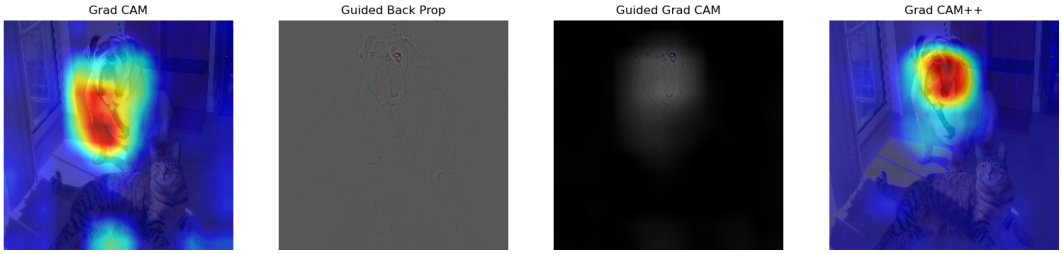

Figure 1: Exemplary visualization of the implemented methods on VGG-16 for one of the dog classes. Grad-CAM++ focuses on more relevant dog features (face). See Appendix for more examples.

a certain prediction by combining feature maps in the final convolutional layer. However, as the calculation of CAMs poses strict constraints on the network architecture, their technique cannot be applied to most CNNs. In 2017, Selvaraju et al. [1] removed these constraints with Grad-CAM (Gradient-weighted CAM) by using the gradient information to represent the importance of each feature map. Two limitations of CAM explanations remained – namely, that they often fail to capture the entire object and that there is a consistent drop in localization performance when multiple instances of the same class are present. Chattopadhyay et al. [3] addressed them both with Grad-CAM++, where positive pixel-wise gradients are incorporated into the weights.

The method of Integrated Gradients [4] proposes a more thorough inspection of the input. It considers the linear path between a *baseline image* (e.g. full black) and the actual input, and calculates the importance of each spatial location by accumulating the pixel-wise gradients along this path. On the other hand, SHAP (SHapley Additive exPlanations) [5] aims to quantify the contribution of a pixel z_i by taking its average marginal contributions across all possible feature subgroups (Shapley value). The intuition is fairly simple. The features can be seen as agents cooperating in a game of correctly classifying an image. Each agent can cooperate with 0, 1, ..., $|\text{features}| - 1$ other features toward the goal. Then, the importance of each agent can be viewed as their individual contribution to the outcome, averaged over all possible groups. SHAP uses sampling techniques to measure these quantities and return them as explanations.

3 Methods

Grad-CAM [1] is a visual explanation method that can argue for why a network has made a certain prediction for a specific image. Given a pretrained CNN-based model, an image and a class of interest c, it generates a heatmap from the relevant layer's feature map activations by first forward propagating the image and then backpropagating the gradients to the layer of interest. Before backpropagation, the gradients should be set to zero for every class except c. The heatmap is defined as the linear combination of the feature map activations. The weight belonging to a specific feature map A^k is denoted by the neuron importance weight α_k^c:

$$\alpha_k^c = \frac{1}{Z} \sum_i \sum_j \frac{\partial y^c}{\partial A_{ij}^k} \tag{1}$$

where y^c is the class score belonging to c and Z is a normalization factor. Since we are only interested in the features that have a positive influence on c, pixels with negative values can be canceled with ReLU. The heatmap can then be calculated as

$$L^c_{\text{Grad-CAM}} = ReLU\left(\sum_k \alpha_k^c A^k\right) \tag{2}$$

Grad-CAM is a generalization of CAM [2] as it works on any convolutional layer (CAM only worked on the last convolutional layer if it was followed by a single fully connected softmax classification layer). This makes it applicable to a wide range of CNN families and capable of generating heatmaps of different detailedness. Another positive attribute it has is that it does not interfere with the base network's architecture, thus allowing for computational efficiency and adaptability.

Guided Grad-CAM, a method also presented by Selvaraju et al. [1], is a combination of Grad-CAM and Guided Backpropagation [6]. It shows the fine-grained details and the relevant edges in an image

Table 1: Classification and localization errors measured on the ILSVRC-2015 validation dataset. We always used the last convolutional ReLU layer for visualization.

Model	Classification error (%)		Localization error (%)	
	Top-1	Top-5	Top-1	Top 5
AlexNet	44.58 (44.2)	21.69 (20.8)	68.04 (68.3)	56.18 (56.6)
GoogLeNet	32.46 (31.9)	11.82 (11.3)	56.89 (60.09)	45.44 (49.34)
VGG-16	30.94 (30.38)	10.87 (10.89)	55.82 (56.51)	44.82 (46.41)

at the same time as localizing the important areas by overlaying the Grad-CAM heatmap on the image created by Guided Backpropagation. In accordance with this, it can be calculated by taking the element-wise product of the outputs of these two methods.

Lastly, Grad-CAM++ [3] is a proposed improvement of Grad-CAM which applies a ReLU on the gradients $\frac{\partial y^c}{\partial A^k_{i,j}}$ to filter out gradients that have a negative influence on the output class (similarly to Guided Backpropagation). Figure 1 presents examples of images produced by the methods mentioned.

Throughout our experiments, we used VGG-16/VGG-16-BN, AlexNet and GoogLeNet (all three pretrained on ImageNet), DenseNet (pretrained on NHS Chest-X-ray14 [7]) and trained a simple three-layer convolutional network on MNIST (see Section 5). For the reproducibility tasks, we used the ILSVRC 2015 validation dataset [8] that contains 50k images of 1,000 categories and the corresponding bounding boxes. Chest-X-ray14 contains 112,120 X-ray images of 14 + 1 different classes (14 of them representing detectable diseases and one implying "no findings"). As bounding boxes are only available for 984 images, our experiment on medical images was restricted to them. The images were resized to 224×224 and the bounding boxes were modified accordingly.

4 Reproducibility study

Localization ability An intuitive application of Grad-CAM's heatmaps is in localization tasks where we are interested in not only the occurrence but also the location of an object. This task can be approached with bounding boxes; it can be viewed as a supervised regression problem where the label $y = (x_{\min}, y_{\min}, x_{\max}, y_{\max})$ is compared against ground truth bounding boxes. Generating labeled bounding boxes can be costly, especially in fields where expertise is needed (e.g., medical data). Because of this, it can be interesting to use them in a weak localization task where the network is not explicitly trained on bounding boxes. Given an image, we can generate a heatmap and convert it to a binary map by e.g. using a 15% threshold. This binary image will then contain multiple clusters around which bounding boxes may be drawn. We isolate the one with the largest area and compare it with the true bounding box by computing the Jaccard similarity $J(\mathrm{box}_1, \mathrm{box}_2) = \frac{|\mathrm{box}_1 \cap \mathrm{box}_2|}{|\mathrm{box}_1 \cup \mathrm{box}_2|}$, also known as IoU score. We can then regard this as a binary classification problem where (m, n) is the size of $\mathrm{box}_{\mathrm{real}}$ and positive predictions can be counted as:

$$\mathrm{box}_{\mathrm{predicted}} \simeq \mathrm{box}_{\mathrm{real}} \iff J(\mathrm{box}_{\mathrm{predicted}}, \mathrm{box}_{\mathrm{real}}) \geq \min\left(0.5, \frac{m \cdot n}{(m+10)(n+10)}\right) \quad (3)$$

Our results are shown in Table 1. The numbers generally lie within $\pm 1\%$ of those found in the original paper (these are in parentheses in the table). The slight differences can be attributed to the lack of information about the layer that was used (in the case of VGG) and more importantly preprocessing (image rescaling). We rescaled the images to 256×256 before applying a 224×224 center crop as this is standard procedure for ImageNet. Overall, Grad-CAM provided fairly impressive localization results, considering that the model was not explicitly trained for this task.

Pointing game Pointing game is another technique for investigating Grad-CAM's localization ability. Originally introduced by Zhang et al. [9], this method extracts the maximally activated point from a heatmap and checks whether it is within the bounding box of the target object category – which, in this case, is a ground truth label. The localization accuracy is then defined as $Acc = \frac{\#\mathrm{Hits}}{\#\mathrm{Hits}+\#\mathrm{Misses}}$ where a point within the bounding box is counted as a hit. In the Grad-CAM paper [1], this metric – henceforth referred to as recall – is extended with the fact that now the top-5 predictions are used

Table 2: Localisation accuracy and recall, measured on the ILSVRC-15 validation dataset [8]. As indicated in the paper, Grad-CAM performs exceptionally well in this context. Note: the experiments in the original paper were conducted on COCO, which contains more fine-grained object boundaries (detailed masks instead of bounding boxes). Their slightly lower result with GoogLeNet (0.7058) can be attributed to the increased difficulty of the dataset [9].

Model	Accuracy	Recall
VGG-16	0.86420	0.82565
GoogLeNet	0.90419	0.76685

Table 3: Rank correlation with occlusion measured on 120 images of the ILSVRC-15 validation dataset. The occlusion maps were generated with patches of size 45×45.

Explanation method	Mean	Standard deviation	Paper on PASCAL 2007
Guided Backpropagation	0.0095	0.1240	0.168
Grad-CAM	0.0505	0.1283	0.254
Guided Grad-CAM	0.0526	0.1287	0.261

instead of the ground truth labels and there is an additional option to reject any of them if the maximally activated point is below a certain threshold. Since the original paper offered no insight into the threshold value, we empirically decided to use 0.5 which is a standard mid-range value in most cases (we noticed that max activations typically ranged from 0.3 to 0.7). Our results are shown in Table 2; for completeness, we chose to showcase both accuracy and recall. They illustrate that GoogLeNet visualizations are good at identifying true positives at the expense of false positives, as indicated by the recall (GoogLeNet visualizations of absent objects are more overconfident).

Image occlusion Another interesting property is how our heatmaps relate to other traditional visualization metrics like occlusion maps. Rank correlation between occlusion maps and heatmaps is a metric that can be used for evaluating the visualization capabilities of Grad-CAM and Guided Grad-CAM. Occlusion maps [10] provide explanations with high local faithfulness. They can be created by using rectangular patches of a predefined color (commonly gray) to mask different parts of an image and then filling the map with the CNN scores of the modified images in areas corresponding to the masked regions. The exact algorithm and the parameters used are not mentioned in the Grad-CAM paper [1] which made this experiment hard to reproduce. We have tested multiple implementations; in the final version, we shifted a rectangular mask along the images with a stride of 1 and padded the original images with half of the patch size. This resulted in images similar to the ones the authors of Grad-CAM provided.

The type of rank correlation was not specified in the Grad-CAM paper, therefore we decided to use the popular *Spearman's rank correlation coefficient*. To calculate this, the maps to be compared need to be downsampled to 14×14, flattened and ranked based on pixel intensity [11]. Due to time limitations (each image requires $224 * 224 = 50176$ passes), we averaged the rank correlations over the first 120 images of the ILSVRC-15 validation dataset. Our results are shown in Table 3, and some examples can be seen in Appendix B.

Overall, we have failed to reproduce the original results of this experiment. Our correlation coefficients are much smaller, which could be attributed to the multiple unspecified hyper-parameters (e.g. patch size, stride, resizing factor) and to the fact that we used a different dataset. We remark that when compared to each other, the relative performance of the methods are in line with the original results, but the coefficients are not big enough to lend credible support to the original values.

User study Similarly to the original paper [1], we conducted a survey in order to compare the trustworthiness of Guided Grad-CAM and Guided Backpropagation [6] using VGG-16 and AlexNet, leveraging the fact that the former is known to be more accurate.[2] The aim of the experiment was to determine whether Guided Grad-CAM is better at showing this claim than Guided Backprop. To do

[2]We note that the original experiment used the PASCAL VOC dataset, however, we decided to use ImageNet to avoid having to retrain the networks.

this, we picked 13 images from the ILSVRC-15 val dataset – images for which both VGG-16 and AlexNet made the same prediction as the ground truth – and generated explanations for them using the two methods. We asked 36 people to rate multiple image pairs; the responses were then mapped to numbers between -2 and 2 in the same way as in the original paper. The summary of the scores is shown in Table 4.

We found that the difference between the two models was not as sharp as in the original paper in either of the two cases, meaning that picking the "better" model is not always as straightforward as the authors claim. Moreover, Guided Grad-CAM achieved a lower average score and a higher standard deviation than Guided Backpropagation, which indicates that it made people more divided on which model is more reliable. Our assumption is that this is because Guided Grad-CAM uses localization to hide certain image regions which can make the explanations more diverse, while Guided Backpropagation produces more uniform visualizations and gives away AlexNet pretty quickly, which, in lack of other factors, made people favor VGG-16 more often. Overall, we feel that this study in itself is not robust enough. The paper does not ensure double blinding or control the relevant background variables, and the results have no statistical significance. In order to draw definitive conclusions, a better evaluation metric would have to be proposed. For more details on this study, please refer to Appendix D.

Table 4: Relative reliability measured with a user study using images of the ILSVRC-15 val dataset.

Explanation method	Mean	Standard deviation
Guided Grad-CAM	0.2714	1.3731
Guided Backpropagation	0.4359	1.2850

5 Novel experiments

Weakly-supervised localization on medical images Since mistakes in medical imaging tasks can have immense impact on human lives, deep explanation methods will play a crucial part in the adoption of neural networks as they enable human experts to efficiently review the decisions of black-box models. Therefore, as our first novel experiment, we measure the weakly supervised localization accuracy of two pretrained DenseNet-121 models (that we will refer to as models "A" and "B"[3]) on the Chest X-ray14 dataset [7].

We generated the heatmaps for the ground truth class using Grad-CAM and Grad-CAM++ on the last convolutional layer of the networks, and binarized them using 50% of the max value as our (empirically chosen) threshold. The predicted bounding boxes were extracted as described in Section 4, and their IoU scores were used to calculate the localization accuracy on the 984 images for which ground truth bounding boxes are available. We found that for model "A", Grad-CAM and Grad-CAM++ correctly localize the disease in 7.2% and 15.5% of the cases, while for model "B" they achieved 15.6% and 19.6% accuracy, respectively. The latter result is comparable with that of Wang et al. [12] which reached an accuracy of 22.8%. It's worth noting that our threshold value was the same for all images, while their approach evaluated two candidates on each datapoint (which leads to improved performance).

We also evaluated Integrated Gradients (IG) and Gradient SHAP[4] as potential alternatives to CAM-based methods. IG is of particular interest because while its authors applied it on a medical localization task, they only shared a single image sample. As the explanations of these two methods do not induce bounding boxes (their heatmaps are very uneven and focus on individual pixels), we devise an alternative metric that measures the ratio of the pixels in the binarized heatmap (removing values that are smaller than 85% of the max value) that lie in the true bounding box (see Table 5). IG and SHAP outperform both variants of Grad-CAM, which means that they might be competing alternatives in other settings as well. However, we note that this evaluation metric can be considered favorable for IG and SHAP as they tend to produce sparser explanations which lead to higher ratios. For qualitative samples, please refer to Appendix E.

[3]Model "A" is available at this URL, while Model "B" is available at this URL.

[4]`https://captum.ai/api/gradient_shap.html`

Table 5: The average ratio of pixels with >0.85 attribution (on a scale of 0 to 1) that fall into the true bounding box.

	Grad-CAM [1]	Grad-CAM++ [3]	IG [4]	Gradient SHAP [5]
DenseNet-121 "A"	0.1313	0.2537	0.2713	0.2583
DenseNet-121 "B"	0.2068	0.2791	0.4245	0.4131

Table 6: Summary of Grad-CAM heatmap activation values for the adversarial pixels in 76 one-pixel attacks.

Model	Before attack				After attack			
	Mean	Std.	Min.	Max.	Mean	Std.	Min.	Max.
VGG-16	0.0609	0.0553	0.0	0.2450	0.0683	0.0716	0.0	0.2840
GoogLeNet	0.3476	0.1984	0.0264	0.9397	0.3039	0.1875	0.0028	0.9302

Measuring sensitivity with one-pixel attacks One of the main motivation behind Integrated Gradients is that most explanation methods (including gradient-based approaches such as LRP [13] or Grad-CAM) violate the so-called *Sensitivity* property, which is satisfied if and only if *for all cases where changing a single feature can change the network prediction, the explanation method gives nonzero attribution to that feature* [4]. This property is clearly desirable; how severely does Grad-CAM violate it in practice? In order to find out, we generated single-pixel adversarial attacks [14] on ImageNet for VGG-16 and GoogLeNet (see Appendix F for examples). Since this method is computationally intensive, we only ran it on the first 250 validation [8] datapoints, and found 76-76 successful attacks for the two models.

Remarkably, upon inspecting the attribution values of the selected pixel (Table 6), we found that with GoogLeNet, Grad-CAM reliably respects the Sensitivity property. For VGG-16, however, it consistently assigned near zero attributions to the adversarial pixels. These results suggest that Grad-CAM can be a sensitive method in particular scenarios, which is a benefit that has not been explored before, to the best of our knowledge. We hope that in the future we will see similar experiments that empirically investigate the theoretical advantages and disadvantages of deep explanation methods.

Measuring fidelity and contrastivity on multiple different methods Inspired by the work of Pope et al. [15], we decided to test Grad-CAM, Integrated Gradients [4] and SHAP [5] (state of the art visualization techniques) with regard to fidelity and contrastivity. For the purpose of this report, we will narrow our focus to the MNIST and CIFAR-10 datasets. For the former, we trained our own network while for the latter we used a pretrained GoogLeNet.[5]

Both metrics are based on intuitive reasoning. Fidelity measures the importance of the salient features highlighted by the explanation for the classification error. It thus measures the difference (decrease) in classification performance when we mask highlighted features (above a specific threshold). Contrastivity quantifies the differentiation between the visualizations of the positive and negative classes. A method exhibits high contrastivity when its explanation shifts dramatically between classes. Assuming two (binarized) visualizations v_1, v_2, contrastivity is defined as $\frac{d_H(v_1,v_2)}{v_1 \vee v_2}$, where $d_h(.)$ is the Hamming distance. We extend the original definition to a multi-class problem by computing the contrastivity for every pair (positive, negative$_i$) and taking the average.

As both methods require binary heatmaps, they are significantly affected by the selected threshold. A low threshold leads to increased fidelity (as we exclude most of the image) and decreased contrastivity. Furthermore, the intensities across the different methods have different interpretations (e.g. Grad-CAM is always positive, while SHAP and IG can exhibit negative values). Instead of using a fixed threshold, we propose using percentiles, namely picking the 10% or 1% percentile of the positive values.

Our experiments highlight three key features (see Table 7). First, on a deeper network, Grad-CAM exhibits the highest contrastivity (see CIFAR-10). Note that Pope et al. [15] found similar relative

[5]Publicly available at this https URL.

Table 7: Summary of Contrastivity and Fidelity of the three visualization methods on the MNIST and CIFAR-10 datasets. The models achieved 96.85 % and 92.73 % test accuracy, respectively.

Metric	MNIST			CIFAR-10		
	Grad-CAM	SHAP	IG	Grad-CAM	SHAP	IG
Fidelity (1 %)	0.501	0.916	0.929	0.782	0.802	0.762
Contrastivity (1 %)	0.647	0.671	0.663	0.887	0.694	0.729
Fidelity (10 %)	0.470	0.916	0.929	0.770	0.798	0.758
Contrastivity (10 %)	0.700	0.696	0.698	0.911	0.720	0.754

results when comparing Grad-CAM with other methods (on different datasets). Secondly, Grad-CAM is on average slightly more affected by the threshold value, as its values are less skewed and evenly spread (due to the bilinear upsampling). Lastly, Grad-CAM's underperformance on the MNIST dataset indicates the fact that the model is not as reliant on its convolutional layer (note that MNIST is a simple dataset which can be mastered with simple fully connected layers). In a sense, Grad-CAM focuses mostly on the first layer of the network, while the other methods explicitly take into consideration the subsequent FC layers as well. This experiment points toward the following: Grad-CAM might not be as efficient as other methods when there is a huge gap between the last convolution and the classification layer (e.g. a deep LSTM). We believe this is a fruitful experiment for future studies.

6 Conclusion

Our report uncovered two reproducibility issues in the original Grad-CAM paper. Firstly, several crucial hyperparameters were left out of the image occlusion experiment's description, and unfortunately the source code of the experiments was not shared. Secondly, we found the experimental design of the user study to be problematic, and it's difficult to assess the original results without seeing the selected images.

Nevertheless, the rest of our results advocate for Grad-CAM's efficacy in CNNs. We reproduced two experiments that showcase its remarkable weakly supervised localization capabilities, and designed novel exploratory experiments in order to evaluate it in new ways. Surprisingly, we found empirical support for the method's potential to be *sensitive*, and showed that it achieves very similar results (often surpassing) to IG and SHAP on the *contrastivity* metric. However, our results indicate that great network performance did not translate as smoothly to good localization in the more specialized medical dataset. Similarly, fidelity experiments suggest that the method might get outperformed by non-CNN based explanation methods when a large portion of the network is non-convolutional. We believe that future studies should address this hypothesis.

7 Self-Assessment and Challenges

Overall, this project proved especially challenging for multiple reasons. We conducted a large number of experiments, each requiring a diverse setup and data handling. Some of the experiments were computationally expensive, often requiring multiple hours to run (even on Google Cloud and Microsoft Azure), and we faced many hurdles due to the lack of hyper-parameter specifications in the original paper's experiments (especially in the occlusion correlation experiment).

In addition to the techniques in the original paper, we implemented Grad-CAM++ and the Fast Gradient Sign Method from scratch and reused the implementations of one-pixel attacks [6], IG and SHAP. Our three novel experiments address distinct aspects of Grad-CAM with original approaches. In particular, the first experiment fairly compares its localization performance with the other 3 explanation methods; the second experiment justifiably combines two papers [14, 4] and produces interesting results; and the last experiment deeply investigates two novel metrics, as introduced in [15]. We thus believe that our report and findings qualify for an A.

[6] Publicly available at this URL.

References

[1] R. R. Selvaraju et al. “Grad-CAM: Why did you say that? Visual Explanations from Deep Networks via Gradient-based Localization”. In: *CoRR* abs/1610.02391 (2016).

[2] B. Zhou et al. “Learning Deep Features for Discriminative Localization”. In: *CoRR* abs/1512.04150 (2015).

[3] A. Chattopadhyay et al. “Grad-CAM++: Generalized Gradient-based Visual Explanations for Deep Convolutional Networks”. In: *CoRR* abs/1710.11063 (2017).

[4] M. Sundararajan, A. Taly, and Q. Yan. “Axiomatic Attribution for Deep Networks”. In: *CoRR* abs/1703.01365 (2017).

[5] S. Lundberg and S.-I. Lee. “A unified approach to interpreting model predictions”. In: *CoRR* abs/1705.07874 (2017).

[6] J. Springenberg et al. “Striving for Simplicity: The All Convolutional Net”. In: *ICLR (workshop track)*. 2015.

[7] P. Rajpurkar et al. “CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning”. In: *ArXiv* abs/1711.05225 (2017).

[8] J. Deng et al. “ImageNet: A Large-Scale Hierarchical Image Database”. In: *CVPR09*. 2009.

[9] J. Zhang et al. “Top-down Neural Attention by Excitation Backprop”. In: *CoRR* abs/1608.00507 (2016).

[10] M. D. Zeiler and R. Fergus. “Visualizing and Understanding Convolutional Networks”. In: *CoRR* abs/1311.2901 (2013).

[11] A. Das et al. “Human Attention in Visual Question Answering: Do Humans and Deep Networks Look at the Same Regions?” In: *CoRR* abs/1606.03556 (2016).

[12] X. Wang et al. “ChestX-ray8: Hospital-scale Chest X-ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases”. In: *CoRR* abs/1705.02315 (2017).

[13] A. Binder et al. “Layer-wise Relevance Propagation for Neural Networks with Local Renormalization Layers”. In: *CoRR* abs/1604.00825 (2016).

[14] J. Su, D. V. Vargas, and K. Sakurai. “One pixel attack for fooling deep neural networks”. In: *CoRR* abs/1710.08864 (2017).

[15] P. E. Pope et al. “Explainability Methods for Graph Convolutional Neural Networks”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2019.

A Methods

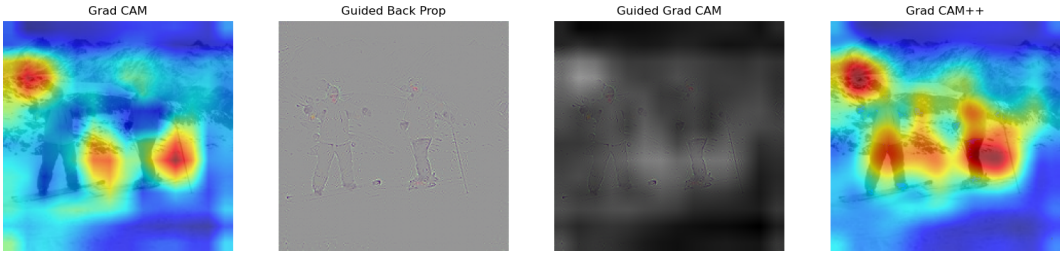

Figure 2: Visualization for *alps*. Grad-CAM++ covers a larger area and includes typical snow equipment. Grad-CAM helps Guided Backpropagation focus less on the humans and more on the background.

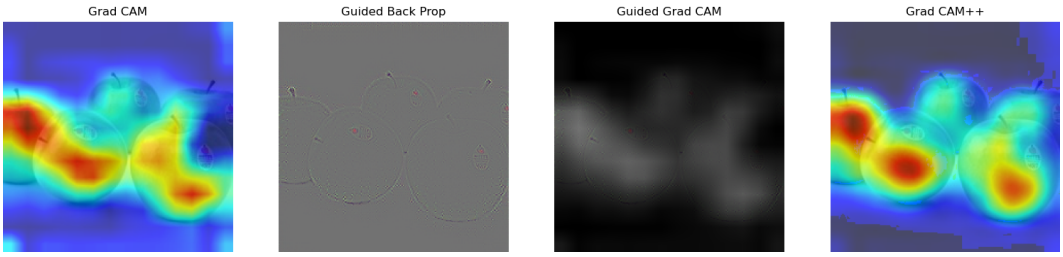

Figure 3: Visualization for *apple*. Grad-CAM++ is able to differentiate between different apples more effectively.

B Occlusion maps

Below are some qualitative results we got by running the Occlusion experiment on the ILSVRC-15 validation dataset. In each row, the first image shows the Grad-CAM heatmap combined with the original image, the second image represents an occlusion map generated using a patch size of 25×25 and the third image is an occlusion map created with a patch size of 45×45. It can be seen that the size of the relevant area increases as the patch size is increased. The occlusion maps' colors range from dark red to dark blue where dark red stands for a specific region's absolute irrelevance in predicting the true label and dark blue indicates complete relevance. What's interesting is that the heatmap in Figure 5 seems to concentrate on the ski pole and the legs, while the occlusion maps focus on the background; similarly, Figure 7 shows that the heatmap is attracted to some pattern-heavy parts of the bowl and slightly to the soup, while the occlusion maps emphasize the entire bowl, especially the edges.

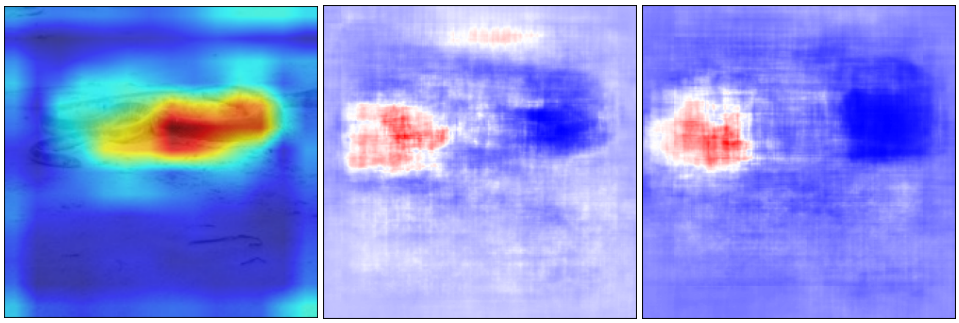

Figure 4: True label: sea snake

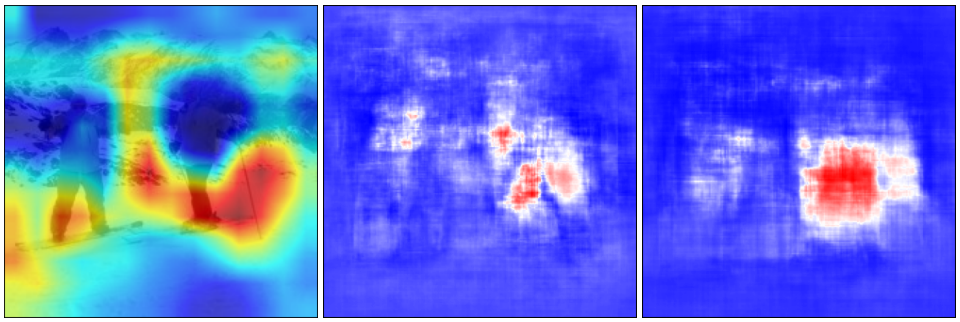

Figure 5: True label: alp

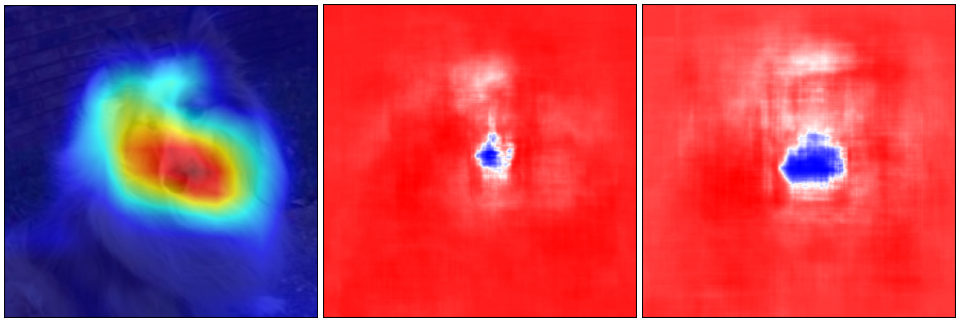

Figure 6: True label: Shetland sheepdog, Shetland sheep dog, Shetland

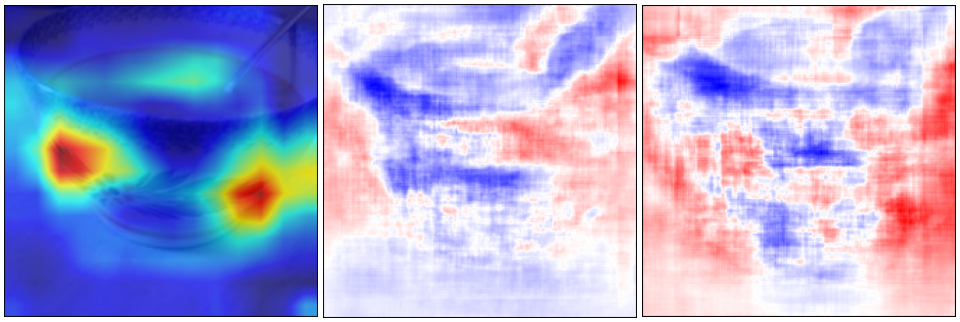

Figure 7: True label: soup bowl

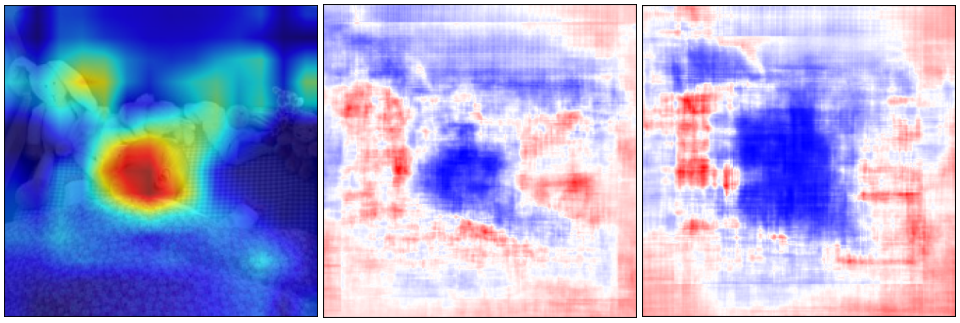

Figure 8: True label: cradle

C Adversarial attacks

According to the original paper, Grad-CAM heatmaps are robust to adversarial noise. In order to verify this claim, we reimplemented the Fast Gradient Sign Method and generated adversarial attacks. Our results successfully confirmed the authors' claim (as shown in Fig. 9). Furthermore, Appendix F demonstrates the same effect on a different type of attack.

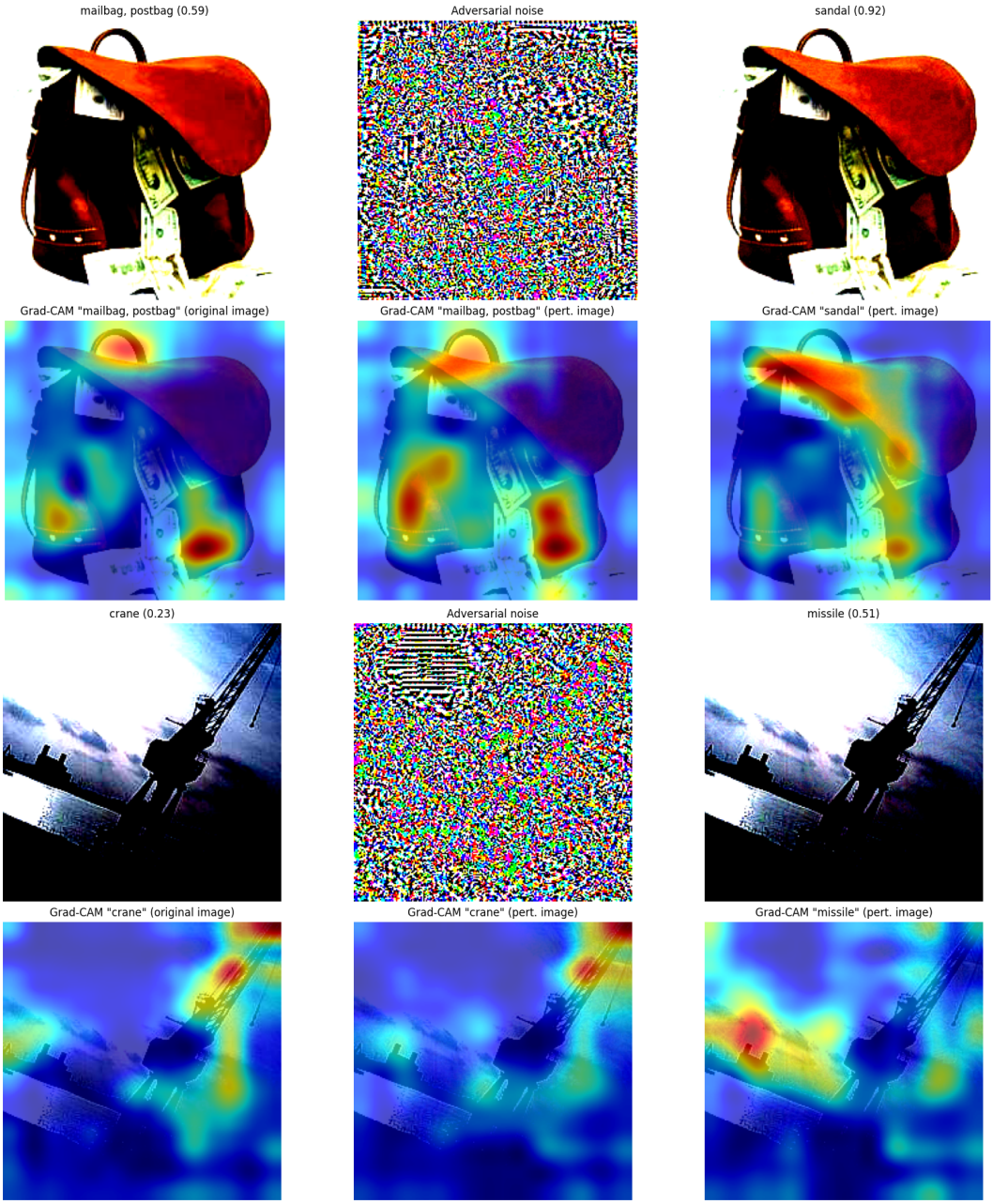

Figure 9: Adversarial attacks, created with FGSM it can be seen that the heatmaps for the ground truth label are mostly unaffected.

D User study

The form for the user study we conducted can be accessed here. What we found while trying to replicate this experiment is that many relevant factors were not addressed in the original paper. Double blinding is not assured as it is unclear whether the authors hand-picked the images in a way that would prove their point or a fairer process was followed (e.g., random sampling). Furthermore, control of relevant factors is not ensured. For instance, AlexNet consistently produces visual artefacts regardless of visualization technique (green patches around the wolf in the images belonging to Agent A in Figures 10 and 12) which affects the final outcome.

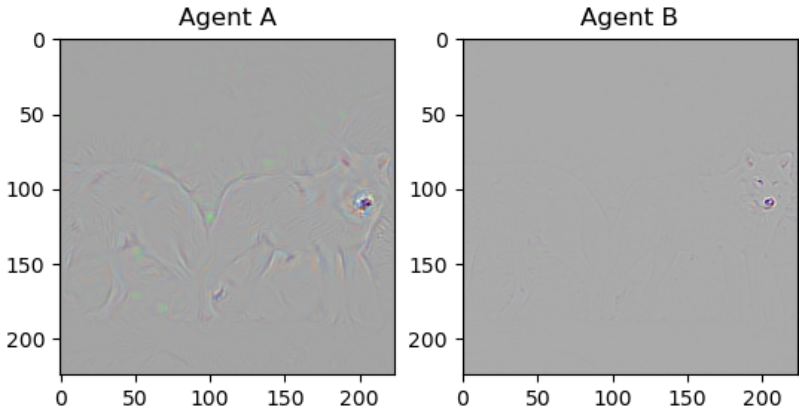

Figure 10: Predicted label: arctic wolf. Explanation method: Guided Backpropagation.

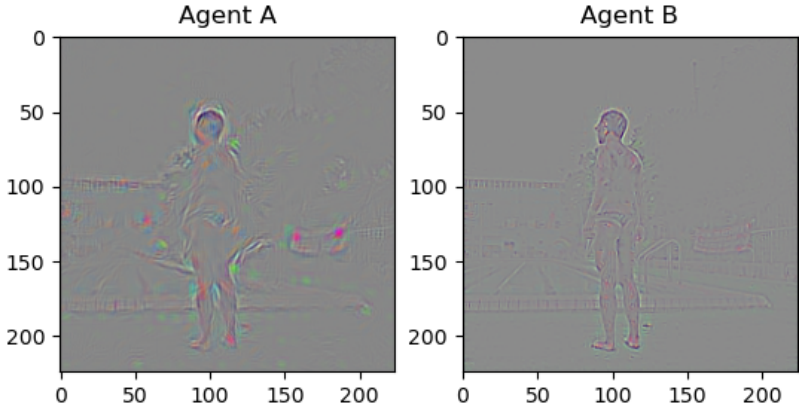

Figure 11: Predicted label: swimming trunks. Explanation method: Guided Backpropagation.

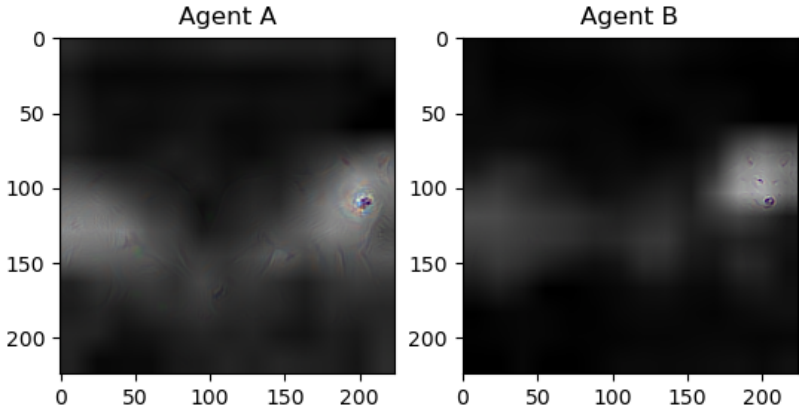

Figure 12: Predicted label: arctic wolf. Explanation method: Guided Grad-CAM.

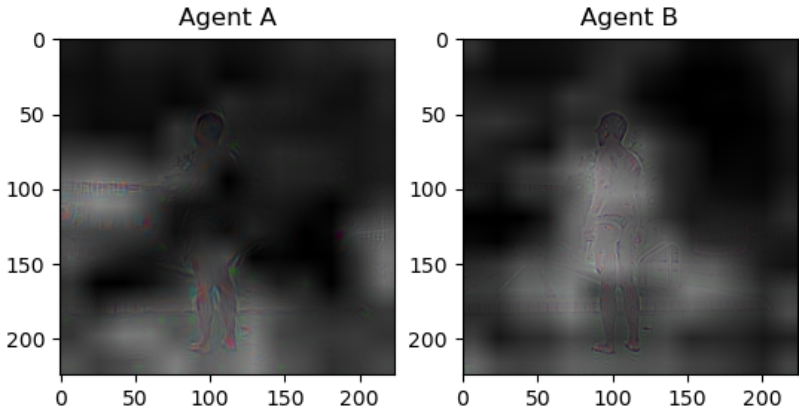

Figure 13: Predicted label: swimming trunks. Explanation method: Guided Grad-CAM.

E Medical imaging

Below we showcase some examples of the heatmaps that were generated in the chest X-ray localization experiment.

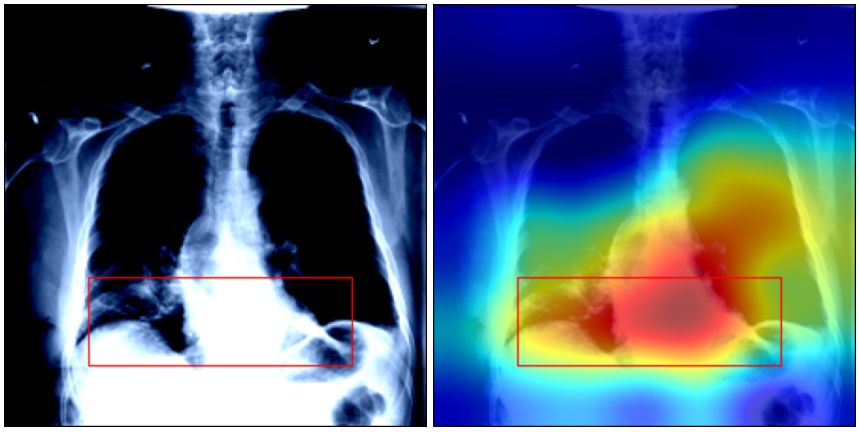

Figure 14: Grad-CAM. Image on the left: original X-ray image with bounding box; image on the right: original image overlaid with the heatmap generated with Grad-CAM.

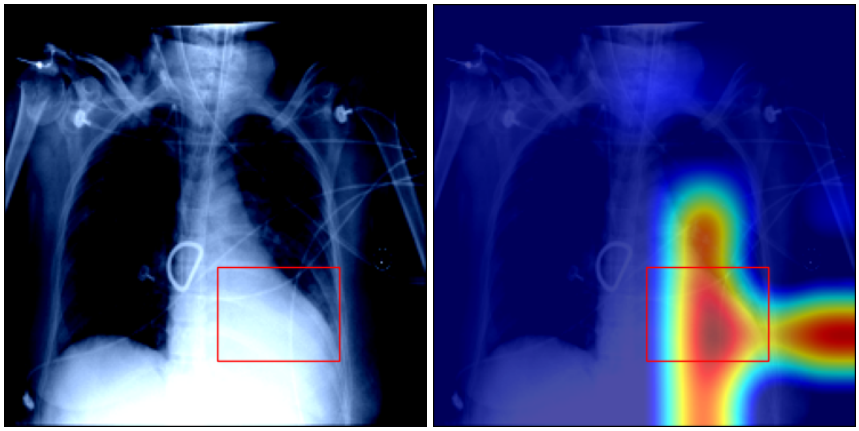

Figure 15: Grad-CAM++. Image on the left: original X-ray image with bounding box; image on the right: original image overlaid with the heatmap generated with Grad-CAM++.

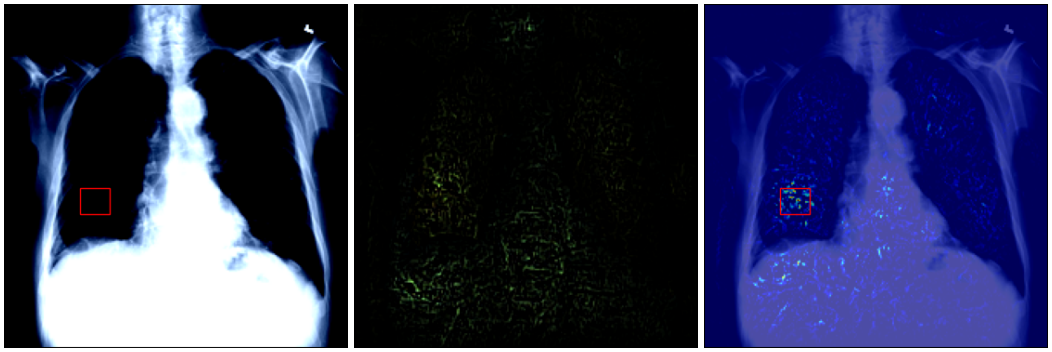

Figure 16: Integrated Gradients. The images from left to right: original X-ray image with bounding box; attribution values; the original image overlaid with the attributions.

Figure 17: SHAP. The images from left to left: original X-ray image with bounding box; attribution values; the original image overlaid with the attributions.

F One-pixel attacks

In each row, the first image shows the heatmap generated with Grad-CAM on the original image, the second image shows the perturbed image where the modified pixel is highlighted with a red circle, and the third image shows the heatmap that belongs to the modified image. In most cases, the maximally activated area in the heatmap shifts to another region.

F.1 VGG-16

Figure 18: Prediction before attack: ballplayer, baseball player. Prediction after attack: assault rifle, assault gun.

Figure 19: Prediction before attack: laptop, laptop computer. Prediction after attack: restaurant, eating house, eating place, eatery.

Figure 20: Prediction before attack: violin, fiddle. Prediction after attack: power drill.

F.2 GoogLeNet

Figure 21: Prediction before attack: trombone. Prediction after attack: red wine.

Figure 22: Prediction before attack: hand-held computer, hand-held microcomputer. Prediction after attack: iPod.

Figure 23: Prediction before attack: menu. Prediction after attack: web site, website, internet site, site.

G Contrastivity and Fidelity

Example of visualizations of the three examined methods: Grad-CAM, SHAP and Integrated Gradients. Due to Grad-CAM's upscaling requirement, its visualizations appear more blurry. Overall, we can see that in both datasets, the visualizations focus on the object present. SHAP differentiates itself from Grad-CAM and Integrated Gradients, as it also focuses on pixels that affect the likelihood conditioned on the rest of the image. For instance, we see that three has highlighted regions in the left portion of it, as the presence of those pixels would transform the three into an 8.

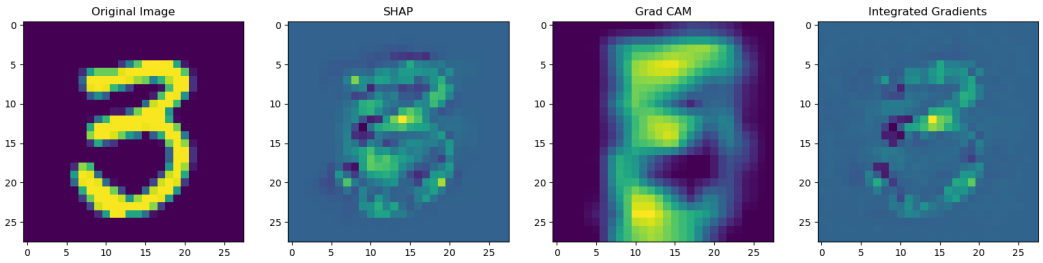

Figure 24: From left to right: Original image, SHAP, Grad-CAM and Integrated gradients. Grad-CAM's visualization was upscaled to 28×28 using bilinear interpolation.