Visualization of Layers Within a Convolutional Neural Network Using Gradient Activation Maps

Dianna McAllister ¹, Mauro Mendez¹, Ariana Bermúdez¹, and Pascal N. Tyrrell ^{*,1,2,3}

ABSTRACT

Introduction: Convolutional neural networks (CNNs) are machine learning tools that have great potential in the field of medical imaging. However, it is often regarded as a “black box” as the process that is used by the machine to acquire a result is not transparent. It would be valuable to find a method to be able to understand how the machine comes to its decision. Therefore, the purpose of this study is to examine how effective gradient-weighted class activation mapping (grad-CAM) visualizations are for certain layers in a CNN-based dental x-ray artifact prediction model.

Methods: To tackle this project, Python code using PyTorch trained a CNN to classify dental plates as unusable or usable depending on the presence of artifacts. Furthermore, Python using PyTorch was also used to overlay grad-CAM visualizations on the given input images for various layers within the model. One image with seventeen different overlays of artifacts was used in this study.

Results: In earlier layers, the model appeared to focus on general features such as lines and edges of the teeth, while in later layers, the model was more interested in detailed aspects of the image. All images that contained artifacts resulted in the model focusing on more detailed areas of the image rather than the artifacts themselves. Whereas the images without artifacts resulted in the model focusing on the visualization of areas that surrounded the teeth.

Discussion and Conclusion: As subsequent layers examined more detailed aspects of the image as shown by the grad-CAM visualizations, they provided better insight into how the model processes information when it is making its classifications. Since all the images with artifacts showed similar trends in the visualizations of the various layers, it provides evidence to suggest that the location and size of the artifact does not affect the model's pattern recognition and image classification.

Keywords: Convolutional neural networks, medical imaging, machine learning

Affiliations

¹Department of Medical Imaging, University of Toronto, Canada

²Department of Statistical Sciences, University of Toronto, Canada

³Institute of Medical Sciences, University of Toronto, Canada

*Please direct correspondence to:

Pascal N. Tyrrell, PhD

Associate Professor

Department of Medical Imaging, University of Toronto

Phone: 416-978-7941 Email: pascal.tyrrell@utoronto.ca

Published online on December 31, 2020

INTRODUCTION

Machine learning is a growing field with numerous uses in technology and science. There is much discussion concerning its applications to a wide variety of fields, notably its applications to the field of medicine¹. It has been said to have the capacity to improve medical care by aiding medical experts¹. One medical field that has been found to benefit from machine learning is medical imaging, because it has been previously shown that it can facilitate image processing and the interpretation of medical images, going so far as to show success in classifying medical images in international competitions. The use of machine learning in medical imaging could greatly assist radiologists in their day-to-day professional tasks and allow for a more precise analysis of the images.

A convolutional neural network (CNN) is one example of machine learning that has been found to be useful in medical imaging. CNNs are machine learning models that attempt to mimic how the brain works by simplifying neurons into units linked in layers to classify visual imagery². Within a CNN, there is a sequence of layers that the input, usually an image of some sort, must pass through before it is classified. Each layer describes the location in the original image for where certain low-level features appear, using the previous layer's output as its input². As the CNN executes more layers, each output represents more and more complex features².

A specific set of layers that repeat multiple times throughout the model exists, where one of the earliest layers is known as the convolutional layer, and is responsible for reducing the input data to create a new image³. The next layer is the normalization layer where whitening is performed, which means that the pixels are normalized to the range $[0,1]$ or $[-1,1]$ depending on the architecture of the model³. Following the normalization layer, there is the Rectified linear unit (ReLU) layer, which functions to obtain the max values of each pixel to provide nonlinearities and to preserve the dimension of the data³. The maxpool layer is the next layer in the sequence and it is used to prevent overfitting, as it reduces the dimensionality of the input and allows for assumptions to be made about features contained in the sub-binned regions³. One of the final layers in the sequence is the average pooling layer, which uses global average pooling of the input to divide it into rectangular pooling regions and to compute the average value of each region to cause down-sampling³. There are various interspersed layers between these specific layers as well, however, they do not have such explicit roles in the model³. This sequence of layers is repeated multiple times to achieve the overall CNN model.

Although machine learning has shown great promise in medical imaging, there is currently a hesitation towards its implementation in all aspects

of medical imaging. This hesitancy is in part due to the inability to understand how the machine attains its final output. This is known as "black box", and it refers to the idea that there is very little understanding surrounding the process through which the machine came to its result or outcome⁴. "Black box" raises ethical concerns of using machine learning to solve medical imaging problems because individuals are incapable of comprehending the decisions made by the machine to reach its final output. This inability to comprehend the decisions made becomes especially concerning if the output of the machine is inaccurate. For example, a study performed by 50 professionals analyzing mammograms with and without computer-aided detection demonstrated a 14.5% decrease in sensitivity for detecting difficult cancers when using the computer-aided detection⁵. These results show that the machine was reaching incorrect outputs for certain mammograms, but it is unknown as to how the model processed these incorrect outputs. Therefore, it would be an unrealistic vision to use computer-aided detection when the harm done could potentially outweigh the benefits. Being able to explain and understand why the machine made the improper decision could propel the innovation of using machine learning in medical imaging because adjustments could be made to the model to provide the correct output and solve the "black box" phenomenon.

There have been previous studies performed where visualization techniques have been used in an attempt to comprehend which features the neural network is observing to reach its output⁷. An example of a visualization technique that can be used is gradient-weighted class activation mapping (grad-CAM). This visualization uses gradients from the previously trained neural network to produce a crude localization map that highlights the important regions of the input image that are used to predict the input image's classification⁶. Nikhil Kasukurthi used grad-CAM in an attempt to understand the erroneous skin cancer detection of his CNN model⁷; which allowed him to realize that the model was paying more attention to the skin color rather than the lesion itself. He was then able to modify his model in order to become more accurate in detecting skin cancer⁷. This is one of many examples where grad-CAM has been proven to be useful in providing insight into how the model reaches its final classification of the given data. This way, it may be possible to alter the model to reduce or eliminate its improper outputs.

Using visualization techniques, such as grad-CAM, at the end of the model has been shown to be useful in determining that seemingly unreasonable predictions made by neural networks have reasonable explanations. However, it still falls short of eliminating the problem as the researcher is uncertain as to which layer or combination of layers are causing the deviation from the accurate output. It is thus important

that research is performed to see if grad-CAM can also be utilized in the earlier layers throughout the model to allow researchers the capability to understand how the model is reaching its final output.

The purpose of the present study is to explore the effectiveness of using grad-CAM to visualize certain layers in a CNN-based dental x-ray artifact prediction model. It is hypothesized that layers deeper within the model when considering a forward pass will be more effectively visualized using grad-CAM, and it can be better used for the interpretability of the model's process. The objectives of this research project are to visualize grad-CAM on various layers interspersed within a CNN model, compare the effect of x-rays with artifacts of different sizes and locations on the visualizations of the layers, and discover if the visualizations provide insight into the interpretability of the details that each individual layer of the model is observing to reach its final classification.

METHODS

This study used a trained dental x-ray binary classification model that classifies whether a given dental x-ray scan was taken on a plate with significant damage or not using a ResNet18 architecture⁷. The data set consists of a total of 2928 x-ray images, where 1404 had no significant artifacts and the other 1524 had significant artifacts⁷. The machine learning model used was developed in Python using the PyTorch machine learning library. Before the model was trained, data augmentation and normalization were performed, and a small amount of training images were visualized to understand the data augmentation. Next, the model was trained, where the amount of change to the model during each step size was scheduled and the model with the smallest loss and highest accuracy was saved. Afterwards, the model was used with a few images to display the predictions before validation occurred. Finally, the pre-trained model was loaded with its final fully connected layer reset, where it was trained and evaluated with the set of validation data. This experiment did not train the model but it used the trained model as input for another python program.

The other Python program also used PyTorch and performed the task of visualizing the activation of various layers of a CNN using grad-CAM⁸. Though there are many techniques that can be chosen to visualize a CNN, grad-CAM was preferable as it requires no retraining of the model and is broadly acceptable to any CNN architecture⁶. The command line arguments that were adjusted were the architecture, the image path, and the weights; all of the other parameters were left as the defaults given by the code⁸. The architecture was altered to match that of the trained model, ResNet18. The image path that was chosen was the file path to an x-ray containing artifacts that was used for validation. Only x-rays that included artifacts contained within

the dataset were used, as it would provide the model with an area of interest to observe and therefore allow for the localization of the grad-CAM. This would also assist in the ability to qualitatively interpret the results of the x-ray overlaid with grad-CAM by the researchers. As the code was executed, all the gradients were set to zero and a forward pass was run on the randomly chosen x-ray. The output would be a grad-CAM based heat map plotted on top of the original x-rays based on the activations.

The target class must also be specified in the previously discussed Python program. The target class represents the specified layer of interest for one iteration of the program. This study used a variety of layers found in the CNN, ensuring the use of layers from the beginning, middle, and end of the CNN to obtain a variation of depths within the model. This would allow for the comparison of visualizations from different points of layers within the model. The layers that were chosen in order from earlier in the model to the end of the model were: Convolutional layer 1 (Conv1), batch normalization layer (BN1), ReLU layer (ReLU), Maxpool layer (Maxpool), various interspersed layers acknowledged as Layer 1, Layer 2, Layer 3, Layer 4, and Average pool layer (AvgPool). For each x-ray used, 9 iterations of the program was performed on the GPU GTX 1080Ti.

This study consisted of seventeen x-rays each containing the same set of teeth with differing overlays of artifacts. Using the same teeth for each image helps to ensure that any differences observed in the visualizations of the images could be attributed to the varying artifacts instead of different images of teeth. Based off of the artifacts, the input images could be grouped into five categories; no artifacts, scratch-like artifacts covering the x-ray, chunky artifacts in the middle of the x-ray, chunky artifacts on the edges of the x-ray, and dot-like artifacts covering the x-ray. The grad-CAM plotted on the input images at the various layers were qualitatively assessed based on localization and intensity of pixels on the heat maps. The three classification of the locations were the location of the artifact, the teeth, and the area surrounding the teeth. The location of the artifact is an important neighborhood to observe as the researchers who created the model are curious as to whether the model is using the artifacts to determine whether the plate is damaged, or if the model is looking elsewhere to make its classification. Analyzing the model's attention to the teeth is also important, as from the dentist's perspective, the ability of the model to detect artifacts on the teeth is more important than those surrounding the teeth in order to prevent dentists from misdiagnosing a dental issue due to an artifact. This assessment was used to evaluate how advantageous using grad-CAM at the specified layer would be to allow the researcher to understand the area of interest the model was observing at that point in the neural network.

Figure 1 The grad-CAM results for image TEST_001_ART_00049_none that contained no artifacts on the x-ray.

RESULTS

The results of one image from each of the five categories can be found above and below, as well as descriptions of what can be observed from each category for the specified layers.

Figure 1 depicts the category of no artifacts on the dental x-ray. For this classification, the grad-CAM visualizations were covering the area surrounding the teeth rather than the teeth themselves. Furthermore, the visualizations of the convolutional layer focus on the outline of the teeth. The BN1 and RELU layers have grad-CAM visualizations covering the complete area around the teeth. The maxpool layer demonstrates a thicker band of visualization around the outline of the teeth and covers less of the area outside of the teeth

compared to the BN1 and RELU layers. Layers 1, 2, 3 and 4 demonstrate blotchy visualizations, highlighting the surrounding area of the teeth. The average pool layer shows no visualization.

Figure 2 has grad-CAM visualizations that begin in the area of the scratch-like artifacts but shift to other locations on the image after later layers are passed. The convolutional layer highlights a large percentage of the teeth area, with little focus on the area of interest. The BN1 and RELU show visualizations on the areas with artifacts, but at a greater intensity than the BN1 and RELU layers. In Layers 1, 2, 3 and 4, the visualizations begin focusing on other areas of the image, specially the regions of the teeth that do not have artifacts. The average pool layer shows no grad-

Figure 2 The grad-CAM results for image TEST_001_ART_00029_severe that contained artifacts that appear as scratches on the upper half of the x-ray.

Figure 3 The grad-CAM results for image TEST_001_ART_00052_severe that contained chunky artifacts nearing the center of the x-ray.

CAM presence.

Figures 3, 4 and 5 have similar trends to that of Figure 2, where the convolutional layer up to Maxpool show visualizations near the artifact and inside of the tooth at increasing intensities. In addition, the Layers 1, 2, 3, and 4 seem to find another area on the tooth to visualize. The average pool layer also shows no visualization.

All of the grad-CAM results from the various x-rays that contained artifacts share commonalities despite having different locations and shapes of artifacts. For each of the x-rays, the convolutional layer has very low intensity and little localization. The batch normalization layer, RELU layer and Maxpool layers have slightly greater intensity and higher localization.

The four Layers have progressive amounts of intensity and localization, until Layer 4 where localization is very precise to areas on the x-ray where the artifact is not present. The average pooling layer shows no intensity or localization. The x-rays that contained no artifacts showed a similar pattern in that as the image was passed deeper the grad-CAM visualization became more specific to one area of the image, but instead the area of interest of the image was usually the area outside of the teeth.

DISCUSSION

The general trend exhibited is that as the x-ray progresses through the layers, the localization and

Figure 4 The grad-CAM results for image TEST_001_ART_00323_moderate that contained chunky artifacts on the edges of the x-ray.

Figure 5 The grad-CAM results for image TEST_001_ART_00075_severe that contained dot-like artifacts on the upper half of the x-ray.

intensity increase to what appears to be the area of interest of the model, whether that be the location of the artifact, the teeth, or the area surrounding the teeth. For the trained CNN to be capable of classifying the images, it uses pattern recognition to compare the features it is observing with those it has already seen in order to classify the image into either the “defective” or “non-defective” category. This means that the CNN also follows a rigorous process, whereby each layer uses the preceding layer’s result as its input. This way, the model can almost entirely ignore features and areas of the image that it has already seen and believes to be unimportant or not of interest for classifying the image.

Table 1 provides detailed insight into the grad-CAM visualizations of each layer for each classification group of images concerning the area of the artifact. As demonstrated by all figures, it was shown that the convolutional layer seemed to have low intensity and localization around the artifact. The convolutional layer is in charge of beginning feature detection by reducing the size and number of pixels to focus in on those that seem to be of greatest importance to the model⁹. A prime example of this is in Figure 3, where the convolutional layer focused in on the bottom edge of the large chunk-like artifact close to the center of the image, meaning that the model believes this to be a feature of interest of the model. The batch normalization layer had similar levels of intensity to the convolutional layer, but the grad-CAM was more localized to the area of the artifact. The batch normalization layer scales pixel values so that they are within a specialized range and therefore novel features are not treated as extremes⁹. The BN1 layer of Figure 2 demonstrates similar levels of intensity to that of the Conv1 layer for the pixels containing the grad-CAM, but there is a slight change in the localization, as the BN1 layer has

fewer areas of interest since the model smoothed out some of the extreme points from its input, which was the output of the Conv1 layer. The rectified linear unit (ReLU) layer seemed to have very similar descriptions to that of the BN1 layer with slightly increased intensity and localization near the artifacts. The ReLU layer plays a key role by replacing all negative pixel values in the feature map by 0, which introduces non-linearity into the CNN⁹. As there are no longer negative pixels, an increase in intensity and localization of the pixel matches the elimination of the negative pixels, which is shown in Figure 4. There were no major changes since the values of the pixels were slightly altered to fit within a range of values, so the input it had received from the BN1 layer’s output is very similar to its own output. The Maxpool layer is summarized to have a high intensity of pixels if there was grad-CAM presence in the area of the artifact. The Maxpool reduces dimensions in the feature map, but retains what the model believes to be the most important features and information that fit a common pattern by keeping the largest pixel value from each area of the feature map⁹. As seen in Figure 5, there is no intensity of pixels or the presence of grad-CAM in the area of the dot-like artifacts, meaning that the model concluded that the areas that contained the artifacts were of no interest when classifying the image, whereas in Figure 2, there is high intensity of pixels in the area of the scratch-like artifacts, meaning that the model recognized that the scratch artifacts were important in classifying the image. Layer 1 showed a decrease in localization of the grad-CAM and similar intensity compared to the Maxpool layer. Layer 1 is the layer in the model that takes in the input shape⁹, which is evident in Figure 3 as the shape of the bottom of the chunk artifact is outlined by the grad-CAM. Layer 2 also takes in the input shape but has around half

Image	Scratch-like	Chunks in center	Chunks on edges	Dot-like
Conv1	<ul style="list-style-type: none"> • Present • Not localized • Low intensity of pixels 	<ul style="list-style-type: none"> • Present • Medium localization • Low intensity of pixels 	<ul style="list-style-type: none"> • Present • Not localized • Medium-high intensity of pixels 	<ul style="list-style-type: none"> • Present • Not localized • Medium intensity of pixels
BN1	<ul style="list-style-type: none"> • Present • Localized • Low intensity of pixels 	<ul style="list-style-type: none"> • Present • Localized • Low intensity of pixels 	<ul style="list-style-type: none"> • Present • Medium localization • Medium intensity of pixels 	<ul style="list-style-type: none"> • Present • Low localization • Low intensity of pixels
RELU	<ul style="list-style-type: none"> • Present • Localized • Medium intensity of pixels 	<ul style="list-style-type: none"> • Present • Medium localization • Medium intensity of pixels 	<ul style="list-style-type: none"> • Present • Medium localization • Medium intensity of pixels 	<ul style="list-style-type: none"> • Present • Low localization • Low intensity of pixels
Maxpool	<ul style="list-style-type: none"> • Present • Localized • High intensity of pixels 	<ul style="list-style-type: none"> • Present • Medium localization • Medium-high intensity of pixels 	<ul style="list-style-type: none"> • Not present • Not localized • No intensity of pixels 	<ul style="list-style-type: none"> • Not present • Not localized • No intensity of pixels near artifact
Layer 1	<ul style="list-style-type: none"> • Present • Not localized • High intensity of pixels 	<ul style="list-style-type: none"> • Present • Medium localization • High intensity of pixels 	<ul style="list-style-type: none"> • Present • Medium localization • Medium intensity of pixels 	<ul style="list-style-type: none"> • Not present • Not localized • No intensity of pixels near artifact
Layer 2	<ul style="list-style-type: none"> • Not present • Not localized • No intensity of pixels near artifact 	<ul style="list-style-type: none"> • Present • Medium localization • High intensity of pixels 	<ul style="list-style-type: none"> • Present • Medium localization • Low intensity of pixels 	<ul style="list-style-type: none"> • Not present • Not localized • No intensity of pixels near artifact
Layer 3	<ul style="list-style-type: none"> • Not present • Not localized • No intensity of pixels near artifact 	<ul style="list-style-type: none"> • Low presence • Little localization • Low intensity of pixels 	<ul style="list-style-type: none"> • Not present • Not localized • No intensity of pixels near artifact 	<ul style="list-style-type: none"> • Not present • Not localized • No intensity of pixels near artifact
Layer 4	<ul style="list-style-type: none"> • Not present • Not localized • No intensity of pixels near artifact 	<ul style="list-style-type: none"> • Not present • Not localized • No intensity of pixels near artifact 	<ul style="list-style-type: none"> • Present • Localized • High intensity of pixels 	<ul style="list-style-type: none"> • Not present • Not localized • No intensity of pixels near artifact
AvgPool	<ul style="list-style-type: none"> • Not present • Not localized • No intensity of pixels near artifact 	<ul style="list-style-type: none"> • Not present • Not localized • No intensity of pixels near artifact 	<ul style="list-style-type: none"> • Not present • Not localized • No intensity of pixels near artifact 	<ul style="list-style-type: none"> • Not present • Not localized • No intensity of pixels near artifact

Table 1 For each category of images that contains artifacts, the presence, localization and intensity of the grad-CAM in reference to the area of the artifacts were summarized.

of the number of nodes compared with Layer 1⁹. In Figure 2, it is shown that between Layer 1 and Layer 2 there is a slight decrease in intensity and an increase in localization since there are less nodes, which means that each node must take on larger amounts of information. Since Layer 2 uses Layer 1's output as its input, it will have a similar visualization but have slightly more averaged out pixels as each node now holds information that multiple individual nodes used to handle. Layer 3 seemed to show no localization or intensity around the artifact for any of the classifications of the artifacts. This layer again takes part in the reduction of the number of nodes, so each nodes holds more information than before. Figures 2-5 each show the loss of localization of the artifacts, meaning that the model is observing elsewhere than the area of the artifacts to come to its classification of each image. Finally, the average pool performs a similar task to that of Maxpool, where it reduces dimensionality, but does so using the average of the values of the pixels of each area of the feature map⁹. All classifications of artifacts showed no localization or intensity whatsoever in this layer. The deduction for the discontinuity in trend by the average pooling layer is due to use of global average pooling. If the last layer feature map has reversed

weights for the adjacent pooling regions (for example, 0.9 vs. 0.1), when global average pooling is performed it could be nullifying the differences in the regions and therefore has weights that are equal for all areas of the image, which would result in no grad-CAM shown, as demonstrated by all figures¹⁰.

When observing the area of the teeth depicted by Table 2, the results for each layer for each classification group of images that contained artifacts were very similar to observing the grad-CAM in the areas of the artifacts as seen in Table 1. The slight difference was that in Layer 3 and Layer 4, there were localization and intensity of pixels shown by grad-CAM on the teeth, whereas when observing the area of the artifact there was very little or no localization and intensity. This could be due to the fact that the model believed that specific areas on the teeth were important to classify each image rather than the artifacts themselves. The results for the images that did not contain artifacts were almost the opposite, where localization and intensity of grad-CAM on the teeth decreased throughout the pass of the model. This can be explained by the fact that perhaps the model used patterns found on the teeth to be able to classify the input image as defective. As there was no pattern in the images that did not contain

	None	Scratch-like	Chunks in center	Chunks on edges	Dot-like
Conv1	<ul style="list-style-type: none"> • Present • Localized • Low intensity of pixels 	<ul style="list-style-type: none"> • Present • Localized • High intensity of pixels 	<ul style="list-style-type: none"> • Present • Localized • Medium intensity of pixels 	<ul style="list-style-type: none"> • Present • Localized • Medium-high intensity of pixels 	<ul style="list-style-type: none"> • Present • Localized • Medium intensity of pixels
BN1	<ul style="list-style-type: none"> • Present • Low localization • High intensity of pixels 	<ul style="list-style-type: none"> • Present • Localized • Low intensity of pixels 	<ul style="list-style-type: none"> • Present • Localized • Low intensity of pixels 	<ul style="list-style-type: none"> • Present • Localized • Low intensity of pixels 	<ul style="list-style-type: none"> • Present • Localized • Low intensity of pixels
RELU	<ul style="list-style-type: none"> • Present • Low localization • High intensity of pixels 	<ul style="list-style-type: none"> • Present • Localized • Medium intensity of pixels 	<ul style="list-style-type: none"> • Present • Localized • Low intensity of pixels 	<ul style="list-style-type: none"> • Present • Localized • Low intensity of pixels 	<ul style="list-style-type: none"> • Present • Localized • Low intensity of pixels
Maxpool	<ul style="list-style-type: none"> • Present • Low localization • Medium intensity of pixels 	<ul style="list-style-type: none"> • Present • Localized • Medium-high intensity of pixels 	<ul style="list-style-type: none"> • Present • Localized • Low intensity of pixels 	<ul style="list-style-type: none"> • Present • Localized • Medium-high intensity of pixels 	<ul style="list-style-type: none"> • Present • Localized • Low intensity of pixels
Layer 1	<ul style="list-style-type: none"> • Present • Low localization • Medium intensity of pixels 	<ul style="list-style-type: none"> • Present • Localized • High intensity of pixels 	<ul style="list-style-type: none"> • Present • Localized • High intensity of pixels 	<ul style="list-style-type: none"> • Present • Localized • Medium-high intensity of pixels 	<ul style="list-style-type: none"> • Present • Localized • Medium-high intensity of pixels
Layer 2	<ul style="list-style-type: none"> • Present • Low localization • Medium-high intensity of pixels 	<ul style="list-style-type: none"> • Present • Localized • Medium intensity of pixels 	<ul style="list-style-type: none"> • Present • Localized • Medium intensity of pixels 	<ul style="list-style-type: none"> • Present • Localized • High intensity of pixels 	<ul style="list-style-type: none"> • Present • Localized • Medium intensity of pixels
Layer 3	<ul style="list-style-type: none"> • Present • Low localization • Medium-high intensity of pixels 	<ul style="list-style-type: none"> • Present • Localized • High intensity of pixels 	<ul style="list-style-type: none"> • Present • Little localization • High intensity of pixels 	<ul style="list-style-type: none"> • Present • Localized • High intensity of pixels 	<ul style="list-style-type: none"> • Present • Localized • High intensity of pixels
Layer 4	<ul style="list-style-type: none"> • Present • Low localization • Medium-high intensity of pixels 	<ul style="list-style-type: none"> • Present • Localized • High intensity of pixels 	<ul style="list-style-type: none"> • Present • Localized • High intensity of pixels 	<ul style="list-style-type: none"> • Present • Localized • High intensity of pixels 	<ul style="list-style-type: none"> • Present • Localized • High intensity of pixels
AvgPool	<ul style="list-style-type: none"> • Not present • Not localized • No intensity of pixels 	<ul style="list-style-type: none"> • Not present • Not localized • No intensity of pixels 	<ul style="list-style-type: none"> • Not present • Not localized • No intensity of pixels 	<ul style="list-style-type: none"> • Not present • Not localized • No intensity of pixels 	<ul style="list-style-type: none"> • Not present • Not localized • No intensity of pixels

Table 2 For all categories of images, the presence, localization and intensity of the grad-CAM in reference to the area of the teeth were summarized.

artifacts and is therefore considered functional, the model needed to search elsewhere, and therefore examined other areas of the image such as outside of the teeth.

Table 3 shows the descriptions of the results for grad-CAM visualization of the area surrounding the teeth. For each classification group of images that contained artifacts, the layers showed almost opposite results than what they showed in Table 2. There was very little if any localization and no intensity of pixels due to the lack of grad-CAM visualizations outside of the teeth. This provides confidence that certain patterns on the teeth were recognized and used to classify the images. For the classification that did not contain any artifacts, such as Figure 1, the localization and intensity of the grad-CAM visualization increased throughout the pass of the model, possibly meaning that as the model looked for the patterns on the teeth and could not recognize them, so it went to search elsewhere, which was around the teeth.

The positioning of the area of interest of each x-ray may not always be where the artifact exists, but is instead the location of interest defined by the model. It is well-known that later layers in a CNN fixate on more detailed and complex aspects of its input rather than

earlier layers, and the results of this study support this notion. Earlier layers seem to be focused on general details of the x-ray that the model is observing, such as curves and lines, meanwhile later layers are attracted to more defined areas which seem to be the areas of interest of the model, such as where the star and C are at the bottom right corner of each of the images.

Near the end of the pass of any image with an artifact, it seemed that the model was no longer looking at the artifact. This provides evidence that the location and size of the artifact causes no change to how well the model is able to classify the damaged plate as one to be discarded.

Observing later layers in the model allows for an easier understanding by researchers of how the model classified the x-ray, meanwhile the earlier layers do not show specifically where the model was analyzing on the x-ray in order to get to its final conclusion. This provides evidence that later layers in CNN models can be more easily interpreted by researchers using them to understand the specific characteristics of each x-ray that the model is using to form its classifications.

The ability to understand what aspect of the image the model is observing could lead to researchers refining their models to decrease the amount of

	None	Scratch-like	Chunks in center	Chunks on edges	Dot-like
Conv1	<ul style="list-style-type: none"> • Present • Localized • Low intensity of pixels 	<ul style="list-style-type: none"> • Not present • Not localized • No intensity of pixels 	<ul style="list-style-type: none"> • Not present • Not localized • No intensity of pixels 	<ul style="list-style-type: none"> • Not present • Not localized • No intensity of pixels 	<ul style="list-style-type: none"> • Not present • Not localized • No intensity of pixels
BN1	<ul style="list-style-type: none"> • Present • Low localization • High intensity of pixels 	<ul style="list-style-type: none"> • Not present • Not localized • No intensity of pixels 	<ul style="list-style-type: none"> • Not present • Not localized • No intensity of pixels 	<ul style="list-style-type: none"> • Not present • Not localized • No intensity of pixels 	<ul style="list-style-type: none"> • Not present • Not localized • No intensity of pixels
RELU	<ul style="list-style-type: none"> • Present • Low localization • High intensity of pixels 	<ul style="list-style-type: none"> • Not present • Not localized • No intensity of pixels 	<ul style="list-style-type: none"> • Not present • Not localized • No intensity of pixels 	<ul style="list-style-type: none"> • Not present • Not localized • No intensity of pixels 	<ul style="list-style-type: none"> • Not present • Not localized • No intensity of pixels
Maxpool	<ul style="list-style-type: none"> • Present • Low localization • Medium intensity of pixels 	<ul style="list-style-type: none"> • Not present • Not localized • No intensity of pixels 	<ul style="list-style-type: none"> • Not present • Not localized • No intensity of pixels 	<ul style="list-style-type: none"> • Not present • Not localized • No intensity of pixels 	<ul style="list-style-type: none"> • Not present • Not localized • No intensity of pixels
Layer 1	<ul style="list-style-type: none"> • Present • Low localization • Medium intensity of pixels 	<ul style="list-style-type: none"> • Not present • Not localized • No intensity of pixels 	<ul style="list-style-type: none"> • Not present • Not localized • No intensity of pixels 	<ul style="list-style-type: none"> • Not present • Not localized • No intensity of pixels 	<ul style="list-style-type: none"> • Not present • Not localized • No intensity of pixels
Layer 2	<ul style="list-style-type: none"> • Present • Low localization • Medium-high intensity of pixels 	<ul style="list-style-type: none"> • Not present • Not localized • No intensity of pixels 	<ul style="list-style-type: none"> • Not present • Not localized • No intensity of pixels 	<ul style="list-style-type: none"> • Not present • Not localized • No intensity of pixels 	<ul style="list-style-type: none"> • Not present • Not localized • No intensity of pixels
Layer 3	<ul style="list-style-type: none"> • Present • Low localization • Medium-high intensity of pixels 	<ul style="list-style-type: none"> • Not present • Not localized • No intensity of pixels 	<ul style="list-style-type: none"> • Present • Little localization • High intensity of pixels 	<ul style="list-style-type: none"> • Not present • Not localized • No intensity of pixels 	<ul style="list-style-type: none"> • Not present • Not localized • No intensity of pixels
Layer 4	<ul style="list-style-type: none"> • Present • Low localization • Medium-high intensity of pixels 	<ul style="list-style-type: none"> • Not present • Not localized • No intensity of pixels 	<ul style="list-style-type: none"> • Not present • Not localized • No intensity of pixels 	<ul style="list-style-type: none"> • Not present • Not localized • No intensity of pixels 	<ul style="list-style-type: none"> • Not present • Not localized • No intensity of pixels
AvgPool	<ul style="list-style-type: none"> • Not present • Not localized • No intensity of pixels 	<ul style="list-style-type: none"> • Not present • Not localized • No intensity of pixels 	<ul style="list-style-type: none"> • Not present • Not localized • No intensity of pixels 	<ul style="list-style-type: none"> • Not present • Not localized • No intensity of pixels 	<ul style="list-style-type: none"> • Not present • Not localized • No intensity of pixels

Table 3 For all categories of images, the presence, localization and intensity of the grad-CAM in reference to the area surrounding the teeth were summarized.

improper classifications made. A further related study that can be performed should analyze if there are specific layers near the end of the model that are best visualized with grad-CAM for interpretability by the researchers. Another study would be to analyze the repercussions caused by human interference with the model after understanding what the model is using to classify the input, as it is possible that after being able to interpret the model, researchers refine their model in such a way that it causes overfitting, which makes the model perform even more poorly on newer datasets than before any alterations were made to the “black box” CNN.

CONCLUSION

In conclusion, this study suggests that later layers would be more beneficial in allowing researchers to understand what areas and details of the input are being used by the model to reach its classification, regardless of where the artifact on the image was located.

ACKNOWLEDGEMENTS

I would like to thank Dr. Pascal Tyrrell for allowing me to conduct this ROP299 research project in his lab, as well as providing guidance on my research. I would also like to thank Ariana Bermúdez for allowing me to use her pretrained CNN model in my research and for helping me to build the code used for my grad-CAM classifications. Finally, I would like to thank Mauro Mendez for his help in debugging the code whenever I could not find the error and for answering all my questions relating to machine learning concepts.

COMPETING INTERESTS

No competing interests declared.

Received: March 31, 2020

Accepted: October 4, 2020

Published online: December 31, 2020

REFERENCES

1. Evgeniou, T. & Pontil, M. in Machine Learning and Its Applications: Advanced Lectures (eds Georgios Paliouras, Vangelis Karkaletsis, & Constantine D. Spyropoulos) 249-257 (Springer Berlin Heidelberg, 2001).
2. Prabhu, R. Understanding of Convolutional Neural Network (CNN) — Deep Learning, <<https://medium.com/@RaghavPrabhu/understanding-of-convolutional-neural-network-cnn-deep-learning-99760835f148>> (2018).
3. Deshpande, A. A Beginner's Guide To Understanding Convolutional Neural Networks, <<https://adeshpande3.github.io/A-Beginner%27s-Guide-To-Understanding-Convolutional-Neural-Networks/>> (2016).
4. Olden, J. D. & Jackson, D. A. Illuminating the “black box”: a randomization approach for understanding variable contributions in artificial neural networks. *Ecological Modelling* 154, 135-150, doi:10.1016/S0304-3800(02)00064-9 (2002).
5. Povyakalo, A. A., Alberdi, E., Strigini, L. & Ayton, P. How to Discriminate between Computer-Aided and Computer-Hindered Decisions: A Case Study in Mammography. *Medical Decision Making* 33, 98-107, doi:10.1177/0272989x12465490 (2013).
6. Selvaraju, R. R. et al. in Proceedings of the IEEE international conference on computer vision. 618-626.
7. McAllister, D. MiDATA, CNN classification for dentistry project, <<https://github.com/diannamcallister/CNN-classification-for-dentistry-project>> (2019).
8. Kasukurthi, N. Visualising CNN Models Using PyTorch*, <<https://software.intel.com/content/www/us/en/develop/articles/visualising-cnn-models-using-pytorch.html>> (2018).
9. Karn, U. An Intuitive Explanation of Convolutional Neural Networks, <<https://ujjwalkarn.me/2016/08/11/intuitive-explanation-convnets/>> (2016).
10. Kukacka, J. Why global average pooling is able to work correctly?, <<https://stats.stackexchange.com/questions/330119/why-global-average-pooling-is-able-to-work-correctly>> (2018).