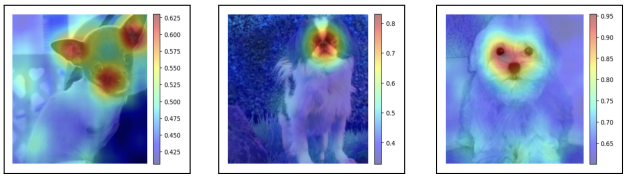
Explainable Image Classifier: case study on Dogs Breeds

Omer Elhussien
African Masters of Machine Intelligence
AIMS-Senegal
oelhussien@aimsammi.org

Abstract

Explainability helps the AI agent through several stages of its life. For instance, when the agent is weak, it can help direct scientists to the weaknesses. The second case is when the agent has the same strength as a specialized person; it increases others' confidence and trust. The last scenario is when the agent is hugely more potent; we could learn from it. In this work, we address the issue of explainability of image classifiers. A dataset of three different types of dogs is used. We used two approaches; one relied on dealing with our models as a complete black box, and the other benefited from the gradient signal to understand what was happening inside the model.

In this work, we are interested to know which parts of the image our model used to make correct predictions about the different classes in the dataset. The images below express our findings with one classifier. We can see how our classifier focuses on specific parts of the dog for each breed to make a correct prediction.



1 Introduction

For the last two decades, marvelous achievements have been seen in Artificial Intelligence (AI). These giant steps are due to several factors, from its connection to several other fields, the growing speed of computational units, and its rigorous reliance on mathematical tools. For instance, trivial AI approaches such as ELIZA rule-based system and fully connected neural networks for image processing tasks have been transformed into Large Language models and convolutional neural networks, respectively. Nevertheless, this progress came with a price, and now scientists are proposing more complex approaches that haunt out-of-the-field users with doubts. AI scientists need to focus more on building tools that could help to increase others' confidence and trust.

Several steps have been taken so far [11, 4, 2, 14, 3, 8, 13, 9, 1] either to design a general framework that could function with various tasks in AI or to build a task-oriented approach that works only for a specific type of structure. In this work, we follow the latter approach. The research problem is to know the key features our model relied on to predict the specific dog breeds in the data. To solve such a problem, we used two approaches from the literature.

The first is Gradient-weighted Class Activation Mapping (Grad-CAM). In this approach, a given image with the desired class passes through the network of interest until the final layer. Then,

the gradients of all other classes are set to zero except for the desired class, which is set to one. Furthermore, the backpropagation journey starts for specific convolutional feature maps. The current approach is seen as one that delves inside the network to understand what is going on. On the contrary, the second approach deals with the given architecture as a complete back-box; only inputs and outputs will be passed through it. This is the case of Randomized Input-Sampling for Explanation (RISE).

In this work, we use six different architectures: AlexNet, ResNet50, ResNet152, ResNeXt101, EfficientNet_ b4, and Vision transformer (vit_ b_ 16). The dataset consists of three dog breed classes: Chihuahua, Japanese spaniel, and Maltese dog[12]. We rely on pre-trained weights for the three classes in the above architectures. During our experiments, we notice exciting patterns such as the network remembering specific parts of the image to make correct predictions, focusing on the correct objects while making wrong predictions, or focusing on other parts of the image. Furthermore, we have realized the same pattern of improvements in these architectures.

The rest of the work is ordered as follows: Section two consists of related work. Then, section three presents the experiments we have tried to address the problem at hand. Finally, we conclude the work with future directions and challenges that faced our research.

2 Related work

The field of eXplainable AI (XAI) is far beyond this work. In contrast, we present some concepts to facilitate our work with image classifiers. ML models are generally classified into two main classes: transparent models, and post-hoc explainability. The former class means models that can be understandable by themselves, such as linear/logistic regression, and decision trees. While the latter class means models that are not interpretable by design, and several approaches are needed to explore them. Some members of this class are Convolutional Neural Networks (CNNs) and Recurrent Neural Networks. Again, for a full explanation, [1] is an excellent review for that.

We could see that our problem is a member of the latter class. It has been known that CNNs could be investigated in two ways. Either by dealing with the CNN as a complete black box and only mapping the output in the input space to see which parts contributed to the prediction; or by delving inside the network and better understanding how the CNN observes the world.

For instance, in [14], they argued the need for explainable tools that pave the way for a better understanding of CNN rather than relying merely on the trial-and-error approach. They proposed a visualization technique that uses a multi-layer DeConvolutional Neural Network (DCNN). It is known that feature maps with strong and soft activations are the main products of an image passed through a CNN. In DCNN, the opposite occurs, and a feature map is passed to reconstruct the maximum activation in the input image, which gives an idea about the parts the network focuses on. They also masked several parts of the input to understand which parts significantly affect the prediction. Moreover, they used their technique to observe the improvement of the training process, and to address challenges that may arise during the training process. Finally, they realized that their models relied on local structures to make predictions rather than remembering some parts of the image. However, their work could visualize only one activation present in a specific layer.

Another different approach could be seen in [3]. They proposed a general framework to tackle the issue of explainability with image classifiers. Their approach relied on the bag of words features with kernel-based classifiers. They used a pixel-wise decomposition to understand the classification process better. They were interested in visualizing the importance of each pixel in the prediction process. These pixels' importance is then presented in a heatmap.

On the contrary, the work of [13] proposed Gradient-weighted Class Activation Mapping (Grad-CAM). Grad-CAM helped to detect bias in the dataset and address failure modes in the models, which could help scientists tackle specific parts of the model. The Grad-CAM approach came as an improvement of CAM[15]. The CNN is modified in CAM by replacing the last fully-connected layers with convolutional layers and average pooling. While in Grad-CAM, the architecture remains the same. Grad-CAM starts with an input image and a desired label passed through the network until a probability is gained for the given label. Then, the gradients for all other classes are zeros except the class of interest, which is set to one. Furthermore, we backpropagate this signal through the network to the desired rectified convolutional feature maps. Finally, these feature maps are combined to produce the Grad-CAM localization.

However, the approach of [9] differs from Grad-CAM. In their work, they considered the CNN a complete black box, limiting their access to the input and output. Randomized Input Sampling for Explanation (RISE) starts with a given image passed through the network, and the predicted probability is recorded. Then, other versions of the same image masked in several parts did the same forward pass and recorded their outputs. These outputs and images are linearly combined- with outputs used as weights for their combination- to produce the final important map. RISE is computationally expensive compared with Grad-CAM due to the massive number of masks. Another challenge RISE may face is the presence of images with several objects of different sizes. In that case, we expect to get noisy importance maps due to the sampling process of masks.

During our implementation, we notice a strange pattern: the CNN focuses on the object of interest but still makes wrong predictions. This type of error could be explained by the work of [6]. In their work, they notice that models trained on ImageNet inherit texture bias from the data, not due to their architecture. This texture bias came with a price, and such models rely mainly on local textures for their predictions rather than object shape. It is known that texture classification is much easier compared with shape recognition. Nevertheless, the former suffers from generalization problems, while the latter is superior with natural images[5].

Finally, we end our literature section with two studies. The first is a case study of skin lesion images[7]. They tailored an existing XAI approach for the problem at hand. For a given image and a classifier, their approach will provide the user with images that get the same or different classification of the input. It also provides the user with a saliency map. The second is a case study of skin cancer[10]. Their framework extracts texture features from the input image after segmentation and clustering techniques. Then, a filtering procedure will follow to ensure confidence and trust. Finally, a hierarchy-based tree is used to interpret the decision process.

3 Experiments

3.1 Grad-CAM technique

In this approach, we rely on the last convolutional layer's gradient signal to better understand the decision process. For a given image to get the class-discriminative localization map of class c, $L^c_{Grad-CAM} \in \mathbb{R}^{w \times h}$, of width w and height h, we need to follow the given steps[13]:

- Compute $\frac{\partial y^c}{\partial A^k}$, where y^c is the score of class c before the softmax, and A^k are the feature maps of the last layer.
- Compute the neuron importance weights $\alpha^c_k = \frac{1}{Z} \sum_i \sum_j \frac{\partial y^c}{\partial A^k_{ij}}$. It captures the importance of feature map k of class c.
- Then, $L^c_{Grad-CAM} = ReLU\left(\sum_k \alpha^c_k A^k\right)$.

Using Grad-CAM, we get interesting results. We present four different scenarios that we encounter during our implementation. It is easy to guess the images on both sides: one is incorrect (to the left), and one is correct (to the right). It is the case when the model looks to the wrong place and makes the wrong prediction; the second case is the model looks to the right place and makes the correct prediction. The second image from the left is an incorrect prediction despite the model looking to the correct position. The last remaining image is a correct prediction. It is the case when the model remembers other parts of the image.

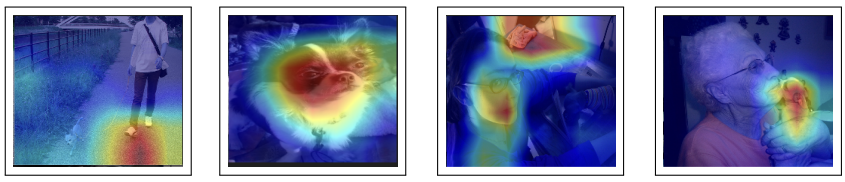


Figure 2: Grad-CAM results.

With this approach, we could not use Vision Transformer since we need access to the last convolutional layer. With our other classifiers, we got the highest accuracy of focusing on the object and making

correct predictions with EfficientNet. However, we have several cases when EfficientNet focuses on a tiny part of the object. From [6], we understand that our network relies on local texture to make its prediction. Also, AlexNet has many cases where it remembers other parts of the image to make a correct prediction.

3.2 RISE technique

In this approach, the main focus is on the generation process of masks. The process is not trivial since independent masking- setting pixels to zero and one- can have adversarial effects, and ample space size. To solve these two issues, we need to rely on bilinear interpolation. We start with more miniature binary masks. Then, we use bilinear upsampling. Finally, we randomly shift the created masks. Using bilinear interpolation benefits our masks by smoothing the edges of the element-wise multiplication between the masks and the given image, and smoothing the importance maps.

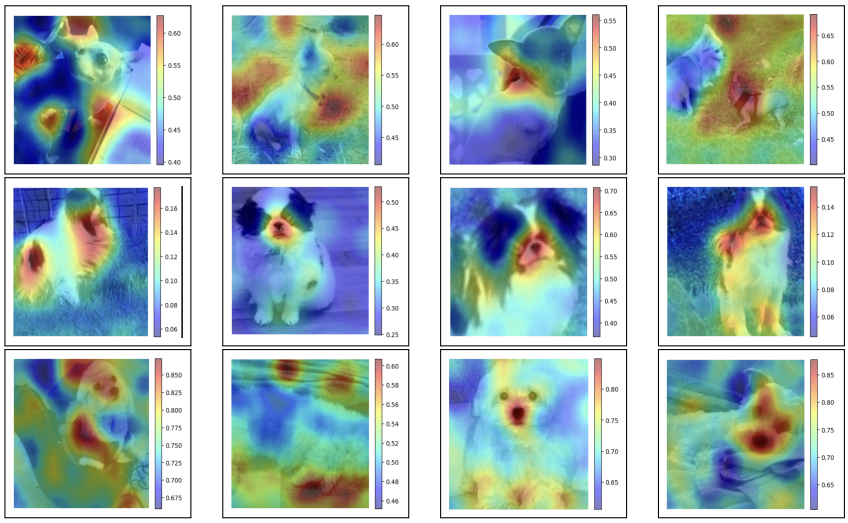


Figure 3: Features AlexNet use to distinguish Chihuahua, Japanese spaniel, and Maltese.

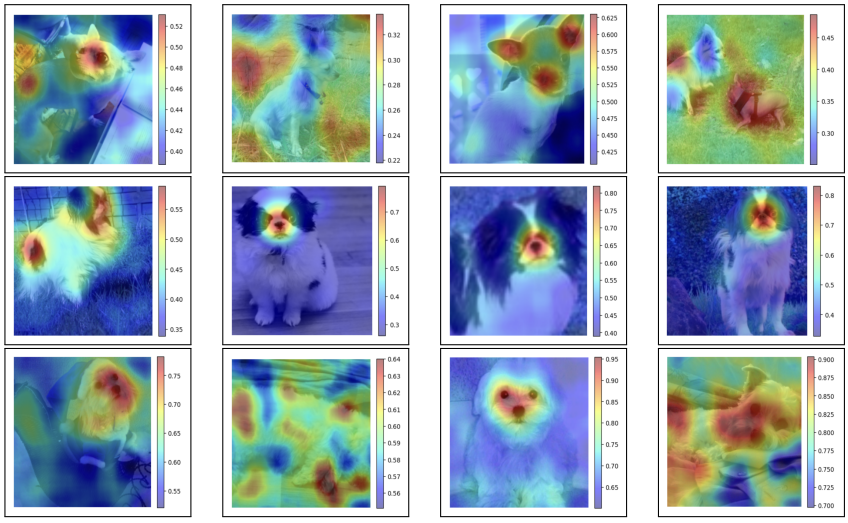


Figure 4: Features ViT use to distinguish Chihuahua, Japanese spaniel, and Maltese.

We will compare the finding of two classifiers, AlexNet, and Vision Transformer. AlexNet, with the Chihuahua class, tries to remember other parts of the image to make correct predictions, while ViT remembers specific features of the same class to reach correct predictions. We see both classifiers

agree on the features they focus on for the Japanese spaniel class. For the last class, we observe that the ViT relies on several parts of the dog when it cannot see its face. The rest of the results can be found in the supplementary materials.

4 Conclusion

In this work, we have used two approaches to explore CNN. We notice that CNN can make correct predictions by looking at the object under question or remembering the general scene. We also observe the texture bias; some networks rely on tiny parts of the object to make correct predictions.

We plan to address some limitations with this work in future work. First, we implemented our approaches to only three classes. We plan to expand our implementation to the one-hundred classes of dogs in the ImageNet dataset. One of the two approaches we relied on considers the CNN a complete black box, and the other benefits from the gradient signal. We want to try an approach that could transform our black box problem into a white box problem. Finally, we are interested in using the two approaches for a training process rather than merely relying on the training and validation error.

5 Supplementary Material

All materials related to this work, such as datasets, results and code, can be found in this repository.

References

[1] A. B. Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. García, S. Gil-López, D. Molina, R. Benjamins, et al. Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information fusion*, 58:82–115, 2020.

[2] M. G. Augasta and T. Kathirvalavakumar. Reverse engineering the neural networks for rule extraction in classification problems. *Neural processing letters*, 35:131–150, 2012.

[3] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, and W. Samek. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one*, 10(7):e0130140, 2015.

[4] N. Barakat and J. Diederich. Eclectic rule-extraction from support vector machines. *International Journal of Computer and Information Engineering*, 2(5):1672–1675, 2008.

[5] K. Burns, L. A. Hendricks, K. Saenko, T. Darrell, and A. Rohrbach. Women also snowboard: Overcoming bias in captioning models. *arXiv preprint arXiv:1803.09797*, 2018.

[6] R. Geirhos, P. Rubisch, C. Michaelis, M. Bethge, F. A. Wichmann, and W. Brendel. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. *arXiv preprint arXiv:1811.12231*, 2018.

[7] C. Metta, A. Beretta, R. Guidotti, Y. Yin, P. Gallinari, S. Rinzivillo, and F. Giannotti. Explainable deep image classifiers for skin lesion diagnosis. *arXiv preprint arXiv:2111.11863*, 2021.

[8] W. J. Murdoch and A. Szlam. Automatic rule extraction from long short term memory networks. *arXiv preprint arXiv:1702.02540*, 2017.

[9] V. Petsiuk, A. Das, and K. Saenko. Rise: Randomized input sampling for explanation of black-box models. *arXiv preprint arXiv:1806.07421*, 2018.

[10] E. Pintelas, M. Liaskos, I. E. Livieris, S. Kotsiantis, and P. Pintelas. A novel explainable image classification framework: Case study on skin cancer and plant disease prediction. *Neural Computing and Applications*, 33(22):15171–15189, 2021.

[11] J. R. Quinlan. Simplifying decision trees. *International journal of man-machine studies*, 27(3):221–234, 1987.

[12] M. Research. Imagenet dogs vs non-dogs dataset. `https://github.com/megvii-research/FSSD_OoD_Detection/issues/1`. Accessed: 2023-08-14.

[13] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.

[14] M. D. Zeiler and R. Fergus. Visualizing and understanding convolutional networks. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I 13*, pages 818–833. Springer, 2014.

[15] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2921–2929, 2016.