PRISM: Principal Image Sections Mapping

Tomasz Szandała[1][0000-0003-4525-0444] and Henryk Maciejewski[1][0000-0002-8405-9987]

[1] Wroclaw University of Science and Technology, Wroclaw, Poland
Tomasz.Szandala@pwr.edu.pl
Henryk.Maciejewski@pwr.edu.pl

Abstract. Rapid progress in machine learning(ML) and artificial intelligence (AI) has brought increased attention to the potential vulnerability and reliability of AI technologies. To counter this issue a multitude of methods has been proposed. Most of them rely on Class Activation Maps (CAMs), which highlight the most important areas in the analyzed image according to the given model. In this paper we propose another look into the problem. Instead of detecting salient areas we aim to identify features that were recognized by the model and compare this insight with other images. Thus giving us information: which parts of the picture were common, which were unique for a given class. Proposed method has been implemented using PyTorch and is publicly available on GitHub: https://github.com/szandala/TorchPRISM.

Keywords: deep learning, xai, convolutional neural networks.

1 Introduction

Deep neural networks show excellent prediction accuracy, but the reasoning for their predictions is often difficult to understand[1]. Moreover it has been proven that classification might be based on latent factors unbeknown to the user. Visualization techniques aim to inspect the model and to prevent reliance on incorrect features. Their task is to identify parts on an input that contributed the most to the output.

Most attribution methods are based on backpropagation of the network's activations from the output back to the input. They are usually a modification of the backpropagation algorithm[2,3,4,5,6] and, for computer vision models, take the form of a saliency map that highlights the decisive regions on the input image.

We are introducing a new method that relies on Principal Component Analysis of features detected by neural network models and interpolates them on the initial image. We call it the Principal Image Sections Mapping (PRISM). The result of the formula is an RGB colored image mask that assigns one color to each feature identified by the model. Moreover the same color will be used to highlight the same feature across all pictures processed in the same batch, of course only if it is present in other samples. This allows exposure of a comparative set of features between images processed in the same batch, and thus facilitates the Explanation by Example technique.

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08751-6_54

In this paper we aim to provide a solution to identify features that contributed to the given classification. During our research, we have performed a deduction experiment based on VGG-16 pretrained model[7]. We take two very similar classes and try to deduce discriminative features between them by clipping the input pictures. Finally we try to generate an adversal attack on the model by swapping identified features between images[8].

2 State-of-the-art

One of the main concerns with the deep networks is that they provide no visual output by themselves, therefore it is not possible to know which part of the image influenced the decision the most[12]. As a consequence, there is a demand for methods that can help visualize or explain the decision-making process of such networks and make them understandable for humans.

Many methods have been devised for this purpose. The first noteworthy group consists of methods that rely on gradient backpropagation. They generate maps of pixels that influence the output of the network.

Other methods can be grouped under that term Class Activation Maps. Here we can find original CAM, Grad-CAM[2] as well several of its derivatives like Grad-CAM++[6], full Grad-CAM[5] and Excitation Backpropagation[4]. The output of these techniques is a heatmap, where the regions that contributed the most for a certain class are highlighted at most. The setback for them is that they are computed for a single chosen convolutional layer. The deeper the layer is, the reliable given saliency map is. However to visualize the obtained map on a referenced image it has to be extrapolated. Since the deepest layers consist of the smallest convolutional masks. the output map might be in a low resolution.

To overcome this obstacle another method has been introduced: Guided Grad-CAM[3]. In it we just multiply the result of Guided Backpropagation and Grad-CAM methods. This gives us pixels that contribute the most to the given class. Likewise any other CAM method can be combined with Guided Backpropagation to determine the most significant pixels.

Apart from these methods several more have been proposed, but our attention caught Explanation by Example. Introduced by Chen et al. in 2019[9] and rated as the most preferred explanation style by the average non-technical end-users[10]. In input domains spanning visual, audio, and sensory data, explanation by nearest training examples offers users an opportunity to compare features across a test input and similarly mapped ground-truth examples. In short: among the training set it points to a set of images that results in a similar response to the examined one. As stated by

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08751-6_54

Jeyakumar et al. it was the most obvious for non-technical consumers, but it is not grounded enough to make deterministic conclusions.

3 Problem description

Our goal was to provide a more faithful method to compare features among similar images. As seen on a picture below, features that determine both wolves and coyotes are their body and head.

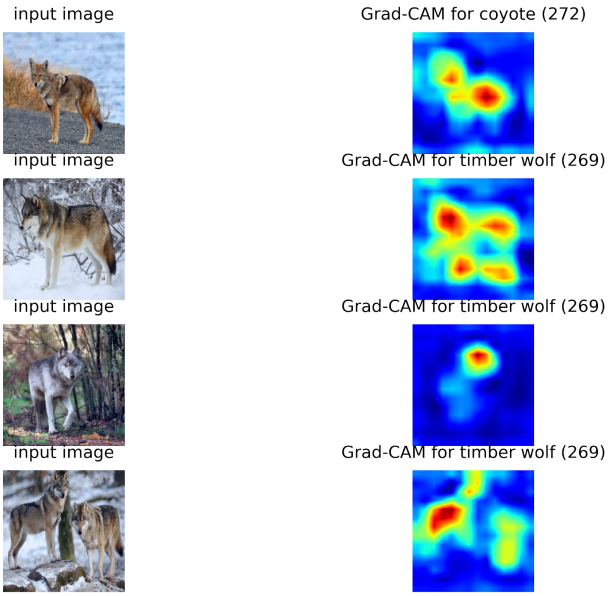

Fig. 1. Grad-CAM output for several pictures

We considered this as insufficient conclusion, therefore we have applied the Guided Backpropagation and combined it with Grad-CAM result which can be seen in figure 2. Results bring us closer to the features that distinguish these two species.
Still the results were disappointing. Despite the fact that we can see features that contributed to the specific class, a human viewer cannot tell which one differentiates between wolf and coyote.

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08751-6_54

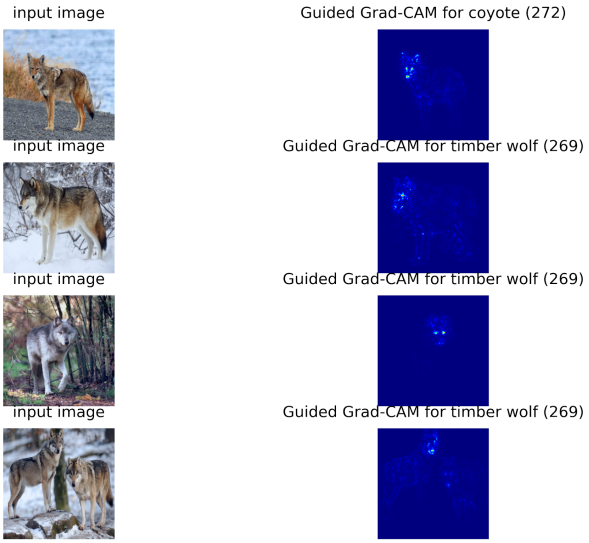

Fig. 2. Guided Grad-CAM result for several pictures

4 Method

We proposed a method that relies on a simple concept: if we look at the final convolutional layer, the Deep Convolutional Neural Networks are only complex representation generators, therefore our focus should be on explaining: what contributed to the given representation vector the most. Here comes Principal Component Analysis, which allows us to reorganize this map into significance-sorted vectors, which could be consequently analyzed further.

PRISM has been designed around this idea. It computes the PCA for the last convolutional layer and truncates the results after the third Principal Component thus receiving an RGB map of features as seen in picture 1.

Assume $\mathbb{A}$ is a set of outputs from each convolutional layer in a network. We choose a single output, preferable from the final layer, but there is no obstacle to study any other layer.. It has a shape of $n \times c \times h \times w$:

n - number of images in a batch
c - number of channels in layer
h,w - height and width of each mask in this layer.
v - multiplication of n*h*w.

Instead of straightforward PCA for dimensionality reduction we use Singular Value Decomposition (SVD) for obtaining Principal Components vectors. Since SVD is applicable only to two dimensional data, we have to reshape the original four dimensional batch into the two dimensional matrix (A') and then center it by subtracting the mean. (eq 1).

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08751-6_54

$$A_{|\mathbb{A}|}^{n\times c\times h\times w} \rightarrow^{reshape} \rightarrow A_{|\mathbb{A}|}^{v\times c} = A'$$

$$A'' = A' - mean(A') \qquad (1)$$

Now we can compute the SVD and then the PCA outcome (eq. 2). Last step is to reshape the obtained matrix back to its original four-dimensional form.

$$U \times S \times V^T = svd(A'') \qquad (2)$$
$$A_{PCA} = U \times S$$

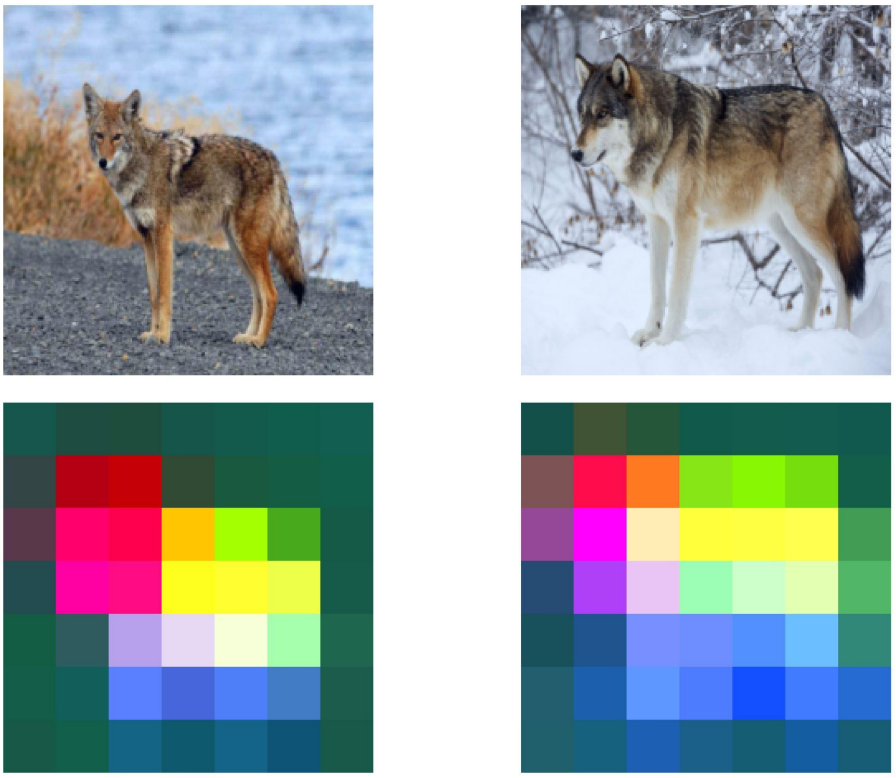

Fig. 3. PRISM raw output for final convolutional VGG-16 depicting coyote and timber wolf class representatives

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08751-6_54

This procedure results in a checkered representation of the processed images (fig. 3.). This Could be further processed using Gradual Extrapolation[11] (fig. 4.) producing a human-recognizable picture with colored features found by the network as seen on figure 5.

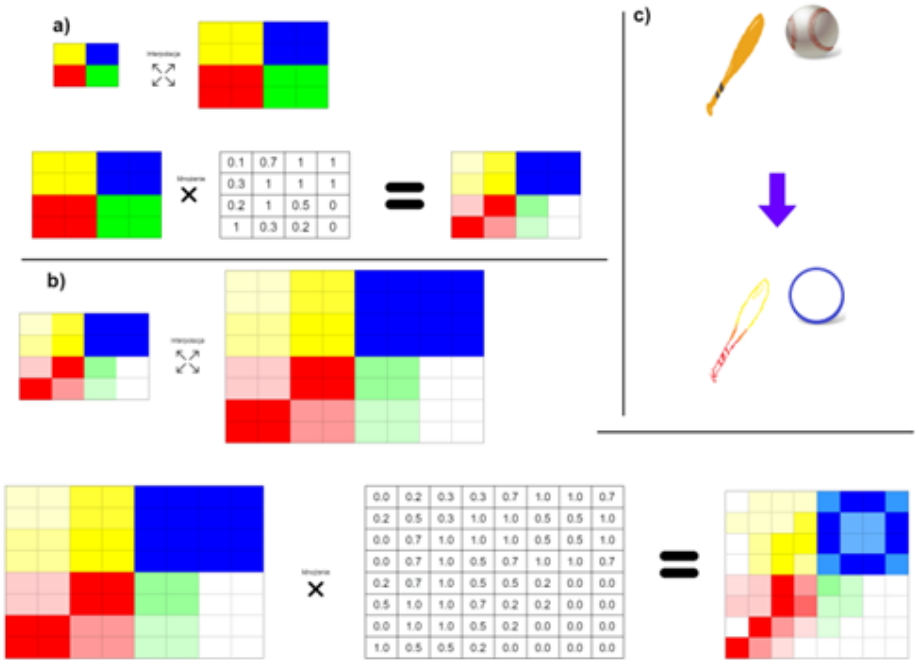

Fig. 4. Schematic concept of gradual extrapolation. Section a) and b) represents Gradual Extrapolation transition from an image of size 2 × 2 to 8 × 8. Section c) shows the concept of transition between image and its restoration using gradual extrapolation

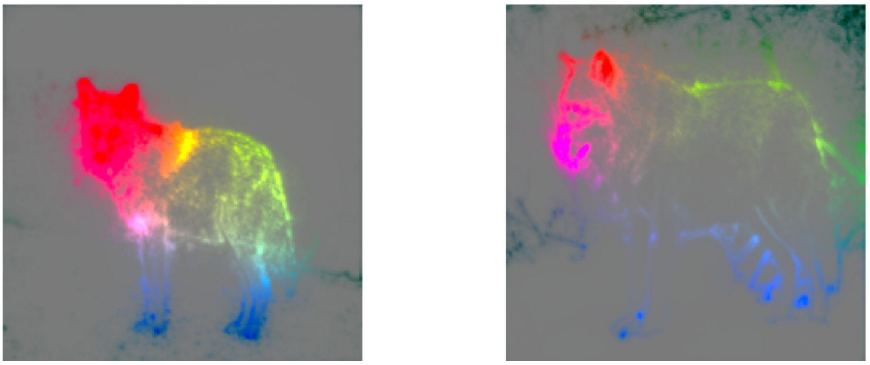

Fig. 5. PRISM output mapped to input image using Gradual Extrapolation technique

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08751-6_54

5 Experiments

In order to prove usability of our method we have performed and analyzed 3 experiments

The first experiment was meant to use PRISM to find and make human-identifiable features that distinguish 2 potentially similar classes. We have started from coyotes and timber-wolves. We have compared these two classes and prepared an example that proves the differentiating features can be identified using PRISM.

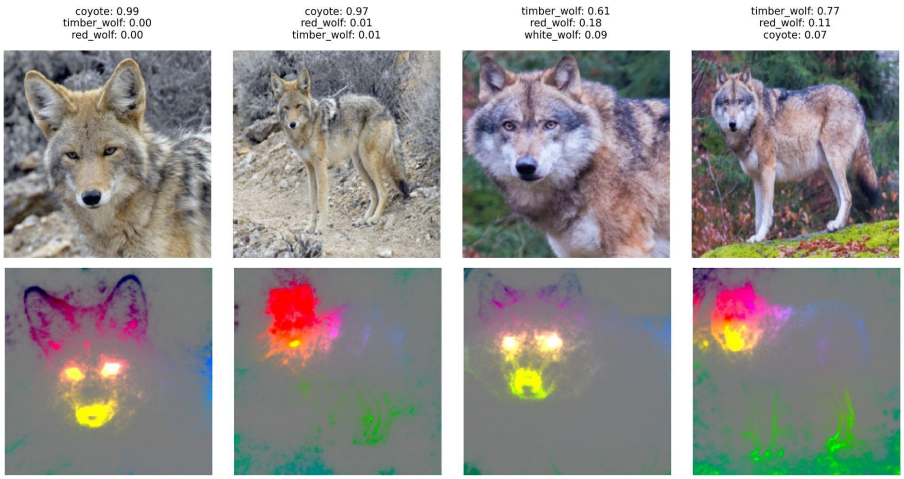

Fig. 6. PRISM output for coyote and its zoomed face as well wolf and its zoomed head

In figure 6 we see PRISM results for 4 pictures. First we started with full bodies of coyote and timber wolf (columns 2nd and 4th). Pawns, tail and back appear to be the similar feature vectors due to similar colors, but the difference appears to be in the animal's heads. So are taking a closer look at them by clipping the original images. Again features identified by PRISM seem mostly equal, but the differences appear in the eyes and ears areas.

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08751-6_54

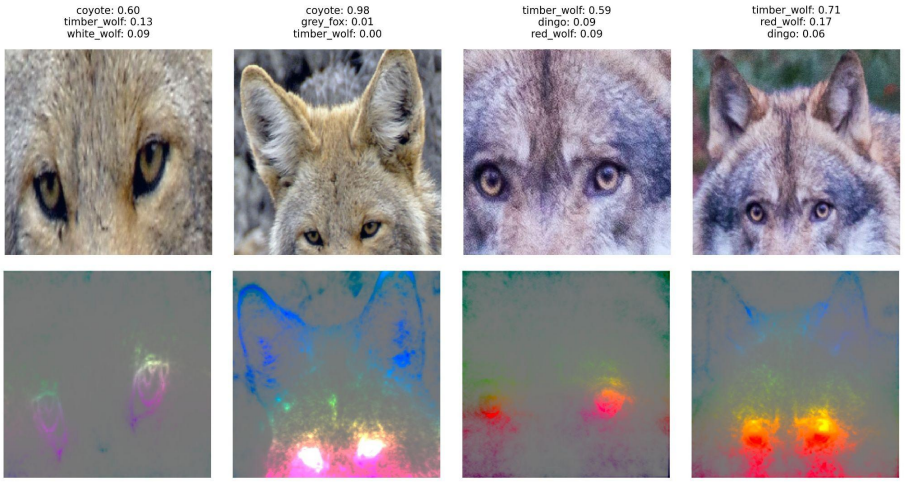

Fig. 7. PRISM output for coyote's and wolf's top parts of the heads and finally eyes

Figure 7. depicts further clipping. First we clip to the top parts of the heads - eyes and ears. This displays similar colors in regards to ears, but different in eyes area, so we perform the final clipping. This leads us to the conclusion that the feature which contributes the most to the distinction between timber wolf and coyote is in their eyes.

In order to check our assumption we have merged coyote eyes into the wolf and other way around. As we see on figure 8. the classification changed significantly. Although the most probable class is still a respective animal, the second guess is the one that we tried to induce. Also note that confidence dropped significantly in favor of the second class. Perhaps better results could be achieved with more sophisticated blending instead of simple paste.

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08751-6_54

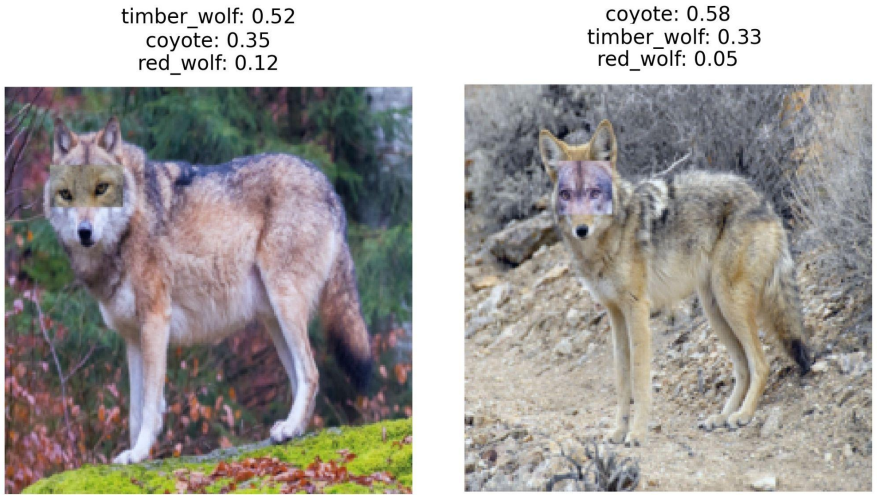

Fig. 8. Classification confidence after swapping eyes between two animals

Of course a question comes, whether basic GradCAM could lead us with a similar deduction path? We have processed the original images and generated GradCAM output for them. As seen in figure 9. the outcome is not easily interpretable, therefore it would be significantly harder to generate a similar adversal attack.

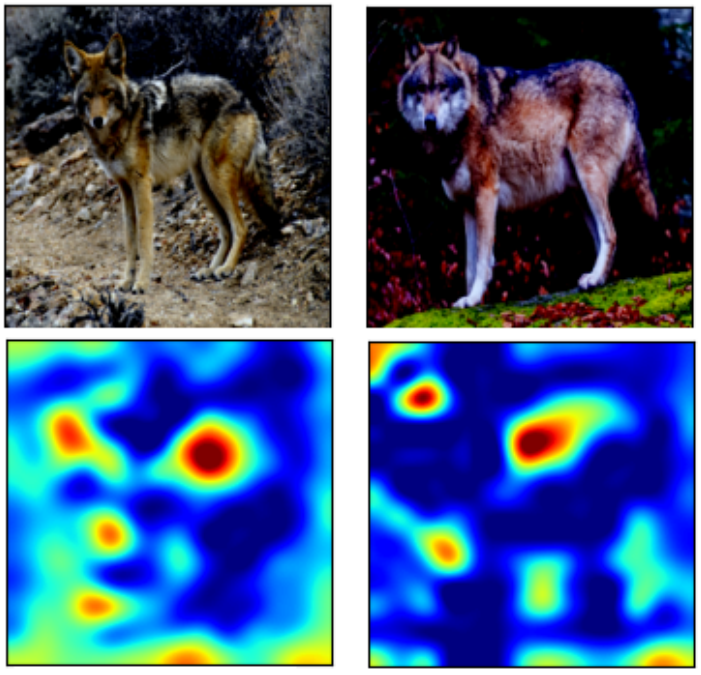

Fig. 9. GradCAM output for original image of coyote and wolf

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08751-6_54

6 Clustering utility

In aforementioned form PRISM is usable mainly for manual inspection of suspicious classes, but it could be used for detecting ambivalent classes if combined with clustering technique, e.g. Self Organizing Maps (SOM). In picture 10 we have presented a first draft of PRISM's clustering utility. We have taken 5 canine classes (color from cluster map in bracket):

- coyote (orange)
- gray fox (red)
- timber wolf (green)
- samoyed (purple)
- border collie (blue)

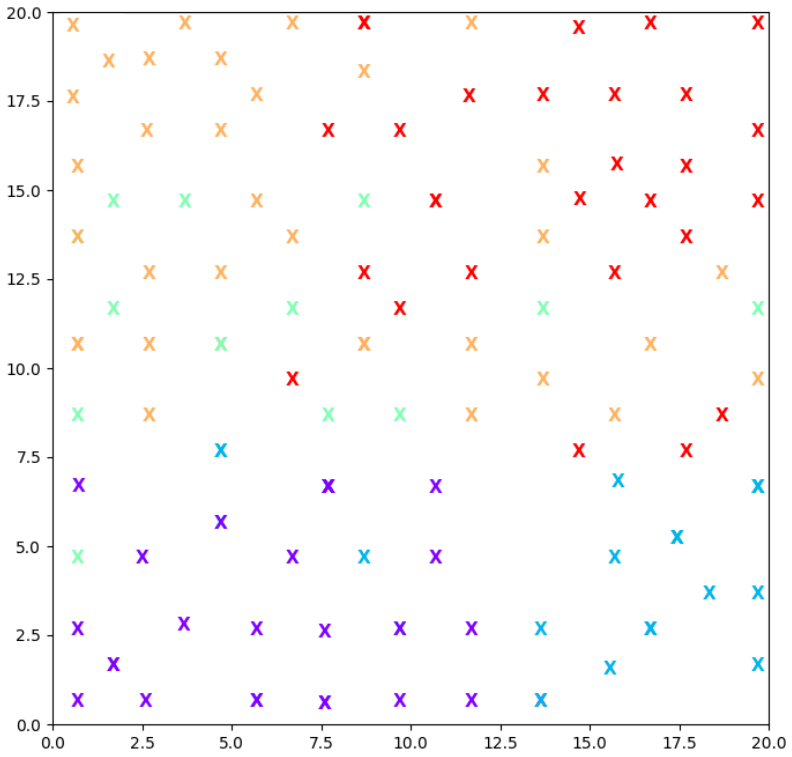

Fig. 10. SOM clustering outcome for 5 arbitrary chosen classes whose representation were generated by PRISM

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08751-6_54

Instead of drawing their PRISM-generated representation we have used them as feature vectors for SOM clustering. Note that PRISM is generating real domain values for coloring, therefore we had to quantize the colors to reduce the amount to a finite number of possible tints.

From figure 10 we can conclude that coyotes could be easily confused with timber wolves and gray foxes. On the other hand the Samoyed and Border collie specimens are well distinguishable from the rest.

7 Conclusions

The proposed method can be a useful tool for CNN examination and can aid in how deep neural networks learn by specifying filters to detect single features, which combined together in the final fully connected layer, provide a classification decision.

Currently the method is dedicated for manual investigation. Our next step will be to improve PRISM's clustering utility, thus giving us an automated detector of problematic classes and possible outliers.

To sum up: PRISM distinguishes from other visualization techniques as it focuses on singular features present on the object. This could give us a multitude of applications, starting from better understanding of the classification process, through automated validation, to pruning of the network.

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08751-6_54

8 Future works

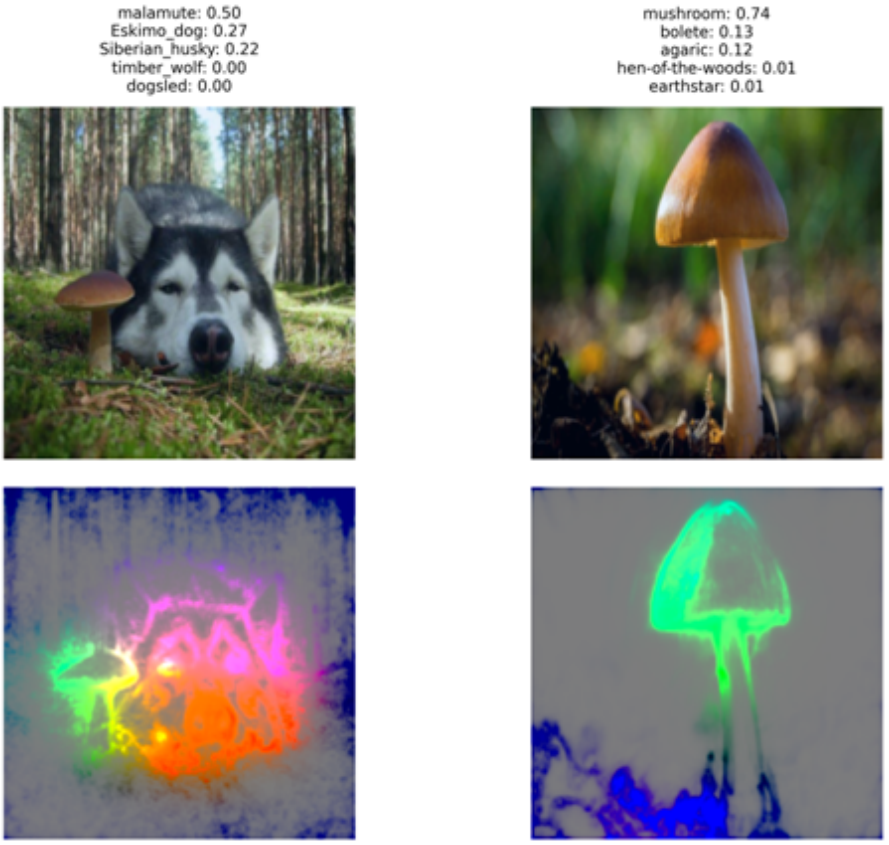

Fig. 11. PRISM output displaying detecting fungus features – a minor issue concerning the proposed method

The biggest setback of the proposed method is lack of actual saliency indicator. In case of an image displaying 2 or more entirely different classes it will still highlight all detected features as seen in figure 11. Left image depicts a dog alongside a mushroom. Lack of saliency factor may suggest that both dog and mushroom are equally important for the network, still if we look at the classification scores we see that model mentions only canine classes to be identified on this image. The countermeasure for this could be blending a saliency method into PRISM's output.

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08751-6_54

References

1. Nguyen, Anh, Jason Yosinski, and Jeff Clune. "Deep neural networks are easily fooled: High confidence predictions for unrecognizable images." Proceedings of the IEEE conference on computer vision and pattern recognition. 2015.
2. Selvaraju, Ramprasaath R., et al. "Grad-cam: Visual explanations from deep networks via gradient-based localization." Proceedings of the IEEE international conference on computer vision. 2017.
3. Bargal, Sarah Adel, et al. "Excitation backprop for RNNs." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018.
4. Zhang, Jianming, et al. "Top-down neural attention by excitation backprop." International Journal of Computer Vision 126.10 (2018): 1084-1102.
5. Morbidelli, Pietro, et al. "Augmented Grad-CAM: Heat-Maps Super Resolution Through Augmentation." ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2020.
6. Chattopadhay, Aditya, et al. "Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks." 2018 IEEE Winter Conference on Applications of Computer Vision (WACV). IEEE, 2018.
7. Simonyan, Karen, and Andrew Zisserman. "Very deep convolutional networks for large-scale image recognition." arXiv preprint arXiv:1409.1556 (2014).
8. Dinh, Vu, and Lam Si Tung Ho. "Consistent feature selection for analytic deep neural networks." arXiv preprint arXiv:2010.08097 (2020).
9. Chen, Chaofan, et al. "This looks like that: deep learning for interpretable image recognition." Advances in neural information processing systems. 2019.
10. Tomsett, Richard, et al. "Sanity checks for saliency metrics." Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 34. No. 04. 2020.
11. Szandała, Tomasz. "Enhancing Deep Neural Network Saliency Visualizations With Gradual Extrapolation." IEEE Access 9 (2021): 95155-95161.
12. Behzadi-Khormouji, Hamed, and HabibRostami. "Fast multi-resolution occlusion: a method for explaining and understanding deep neural networks." Applied Intelligence (2020): 1-25.

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08751-6_54