Problems of identification in information systems

UDC 004.93 **doi:** https://doi.org/10.20998/2522-9052.2025.3.01

Oleksii Gorokhovatskyi, Olena Peredrii, Oleh Teslenko

Simon Kuznets Kharkiv National University of Economics, Kharkiv, Ukraine

MULTIPLE RECURSIVE DIVISION EXPLANATIONS FOR IMAGE CLASSIFICATION PROBLEMS

Abstract. The aim of the research. In this paper, the approach to search for multiple explanations of the CNN image classification case is proposed. **Research results.** The core of the method is recursive division (RD), that performs the perturbation of the input image with hiding different rectangular parts. The explanation is represented as a complementary images pair (CIP): two images that allow us to visualize the parts of the image which are important enough to change the class of the input image when hidden and at the same time are important enough to preserve the initial classification result when visible. The parameters of RD method are discussed to choose the criteria to stop the processing when few explanations are found or the further processing requires too much time and/or memory resources. Two approaches to merge multiple CIP back to single explanation using SLIC segmentation were proposed. They allowed us to reduce the useful image explanation area and sometimes find more visually attractive CIP compared to previous RD implementations. Such merging is not strictly required just multiple CIP explanations are good enough for analysis of the CNN. **Conclusion**. The implementation of the proposed approach for cats and dogs breed classification problem was compared with other popular methods like RISE and Grad-CAM, the benefits and drawbacks are discussed. The performance analysis confirmed the advantage of the proposed methods as they are comparable or faster with known and allow us to find multiple explanation images.

Keywords: Image Classification; CNN; AI; Black-box; Explainability; Image Perturbation; Hiding Parts; Recursive Division (RD); Complementary Image Pair (CIP).

Introduction

Artificial neural networks (ANNs) and convolutional neural networks (CNNs) have established themselves as the successful technology for image classification problems. The effectiveness of these models, along with their ability to automatically learn features from raw pixel data, has played an important role in their widespread usage.

Despite their success in achieving high predictive accuracy, ANNs/CNNs are often characterized as "black-box" models due to the intricate nature of their architecture. This makes it challenging to understand the reasoning of the decisions being made by these models, creating significant trust issues and adoption in domains where the ability to explain and justify decisions is required for building information processing systems, including medical imaging, object detection, autonomous driving, and many others.

In response to the interpretability challenges presented by artificial neural models, the field of Explainable Artificial Intelligence (XAI) has emerged as a branch of research. XAI aims to develop methods and tools that can make the decision-making processes of AI models more transparent and interpretable/explainable to humans.

The fundamental goal of XAI is to bridge the gap between the high predictive power of CNNs and their limited interpretability, enabling humans in some way to understand and trust the decisions made by these neural-networking systems. The ability to understand the reasoning behind a CNN's prediction became a fundamental prerequisite for the widespread and ethical adoption of such models in real-world applications.

Another path that has evolved as a separate field of research is to develop initially interpretable models instead of searching for explanations of existing models.

1 Literature review

A lot of methods aimed to enhance user trust and facilitate the responsible deployment of CNNs applied to image classification problems have been proposed in various researches over the last decade [1–11]. There are classifications of methods into model-agnostic and model-specific (showing whether they are applicable for all existing models or only specific ones), and into local and global methods, that makes it possible to explain the particular output by the model or the entire model behavior. Mostly, the practical usage of these methods relates to the search for the explainability of the particular image following by the interpretation of it by humans because of subjectivity of the entire process [3].

The most famous approaches to find the explanations are SHAP [12], LIME [13], RISE [14, 15], and Grad-CAM [16].

SHAP (SHapley Additive exPlanations) [12] assigns an importance value for a particular prediction for each feature in the input signal. This value represents the contribution of the feature to the difference between the actual prediction and the average prediction across the dataset.

LIME (Local Interpreted Model-Agnostic Explanations) works by approximating the behavior of a black-box model near a specific prediction using a simpler, more interpretable surrogate model. It creates a perturbation image from the input image, for example by selectively turning off its parts. By observing how these perturbations affect the decision of the classifier,

LIME generates a subset of perturbed samples along with their corresponding predictions. Subsequently, the interpretable model is trained on this generated data to mimic the behavior of the initial black-box classifier. The weights assigned to each feature in the resulting local model indicate its relative importance for the model.

One known drawback of the LIME is the instability of the explanations. Due to the random sampling process involved in creating the perturbed data, LIME explanations for the same model and input can vary across different runs. Additionally, LIME's performance is sensitive to the choice of several hyperparameters, e.g., the size of the neighborhood around the instance being explained, the quantity of features (or superpixels for images) used to represent the image. Determining the good values for these parameters can be a non-trivial task and might require experimentation and tuning.

Another known method for explaining CNN predictions in image classification is RISE (Randomized Input Sampling for Explanation) [14, 15]. RISE calculates the importance of each pixel in the input image by generating a large set of random binary masks, and applying each mask to the original input image, occluding different parts of the image. For each masked image, the classification is performed. The importance of a particular pixel is then determined by aggregating the predictions of all the masked images where that pixel was not occluded.

The main drawback of RISE is its computational cost. RISE typically requires a large number of random masks (1000, 2000, etc.) to be generated and stored in memory to obtain a reliable estimate of pixel importance. This extensive sampling process can be particularly time-consuming and computationally expensive for large CNN models and high-resolution input images.

Gradient-weighted Class Activation Mapping (Grad-CAM) [16] is another popular method for generating visual explanations for CNNs classification results. Grad-CAM analyzes the internal gradients of the CNN to create a localization map that highlights the image regions important for predicting a specific class. The process involves first performing a forward pass of the input image through the CNN to obtain the feature maps of the final convolutional layer and the score for the target class. Then, a backward pass is performed to compute the gradient of the target class score with respect to these feature maps. These gradients are globally average-pooled across the spatial dimensions to obtain neuron importance weights for each feature map, indicating their contribution to the target class prediction. Finally, a weighted linear combination of the forward activation maps of the final convolutional layer is computed using these weights, followed by a ReLU activation function, resulting in the Grad-CAM localization map.

The resulting heatmap has the same spatial resolution as the feature maps of the final convolutional layer, which is typically much lower than the input image, leading to a low-resolution explanation. In some instances, due to the gradient averaging step, Grad-CAM might highlight regions that were not actually used by the model for classification, potentially leading to unreliable explanations. Another limitation is that Grad-CAM may struggle to properly localize objects when an image contains multiple instances of the same class.

While the explanation methods mentioned above (as well as numerous other methods based on them) provide valuable insights into the decision-making processes inside CNNs, the reliance on a single explanation image often presents limitations in achieving a complete understanding of the model's reasoning. Typically, an explanation in the form of single image highlights the most dominant features that influenced the classification result but may not capture the influence of all contributing features. Users may believe that the model's decision is solely based on this most prominent feature, even if other factors also contributed significantly to the classification.

Presenting multiple explanations for the same input image and its classification can lead to a more robust interpretation. If different explanation methods consistently highlight similar regions or features as being important, it can increase confidence that these are indeed the key factors driving the model's decision. On the other hand, if different methods emphasize different aspects of the input, it could indicate that the model's reasoning is more complex or that different features contribute to the prediction in various ways. Analyzing these patterns of agreement and disagreement across multiple explanations can provide a better understanding of the model's behavior.

The Structured Attention Graphs (SAG) based on beam search, were introduced [17, 18] to generate multiple explanations in the form of Minimal Sufficient Explanations. These explanations consist of minimally masked images that retain the original classification result with a high probability (>0.9) while preserving essential features. However, a key challenge arises due to the factorial growth in the number of possible masked images, necessitating the division of each image into 49 distinct masks. Furthermore, research has demonstrated that CNNs have more than one way to classify image that justifies the necessity to search more than one explanation.

We have proposed Recursive Division (RD) as a way to search for the explanations for CNN image classifier [19–21] and used it before to generate fast single explanation of the classification result in a form of complementary images pair.

The goals for this paper include:

- to formalize the recursive description of RD, termination criteria and hyperparameters for it;

- to evaluate the possibility to generate multiple explanations of the image classification result as complementary images pairs (CIP);

- to embed superpixels segmentation at the final stages of RD to produce better and more specific visual explanations (instead of producing rectangles only), and build the single explanation result gathered back from multiple ones.

Our contribution in this paper includes the modification of the RD method proposed before in order to find multiple explanations in a form of CIP, improve their visual quality with SLIC segmentation, and reduce the size of the explanatory regions.

2 Complementary images

The main idea we follow across this research is the pair of complementary images (Complementary Images Pair – CIP). We think that the single explanatory image with the parts which are important (that is typical for a lot of other perturbation explanators) is not enough. The motivation for this is that if the part of image is important for the classification, this doesn't guarantee automatically that the other parts are not important enough at the same time.

Let's assume, that we have some black-box image classifier and we used it to predict the class for some initial image, the classification result has the label of class C. Our goal is to find such pair of images (created with perturbations from the initial one), first one of which is still classified as C, but the second one has other class label. Both these images should have different parts hidden similarly, and each part of the initial image may be hidden only in one image in this pair (Fig. 1).

Forming CIP in such a way make us sure that hidden parts are important enough to change the classification result of the CNN, and vice versa – that saving these parts is important enough to preserve the initial classification result.

Fig. 1. Complementary Images Pair (CIP)

3 Multiple Recursive Division Explanations

In our previous researches [19–22] we were focused on the search of the single CIP to deliver explanation result as quick as possible. But the initial version of algorithm proposed in [23] demonstrated the ability to find multiple explanations out of the box.

The entire processing pipeline remained as described in our previous research. Let it be $O = B(I)$ as the output vector that comes from the black-box CNN classifier for the input image I, the length of the vector O is K and corresponds to the quantity of classes being classified. Let C be the label of the classification result and we assume that this classification is correct. We mark the quantity of images to divide image for as w_1 and h_1 for the first division of the initial image, and w_m and h_m for all subsequent ("m" stands for "middle") division layers.

The initial image I is split into $w \times h$ non-intersecting parts, where w is the quantity of horizontal parts, and h is the quantity of vertical ones, $\Delta w = width / w$, $\Delta h = height / h$, $width$, $height$ – are the width and height of the image I respectively. Replacing each generated part with the hiding color (e.g., black) allows to obtain the set of $w \times h$ perturbed images $\{I^{i,j}\}, i = \overline{1,w}, j = \overline{1,h}$:

$$I^{i,j} = I \times M^{i,j},$$

$$M^{i,j} = \begin{cases} 0, \forall (x,y): x \in [(i-1)\Delta w, i\Delta w], y \in [(j-1)\Delta h, j\Delta h], \\ 1, \text{otherwise}. \end{cases}$$

It is worth noting, that selection of color to hide part with could influence classification results itself.

Perturbed images are generated one by one and checked immediately in place that allows us to save memory.

The main idea of RD is shown in Fig. 2. The initial image has "Maine coon" class label and is classified correctly. At the first division stage with $h_1 = w_1 = 2$, four perturbed images are generated and classified immediately. All of them are still "Maine coon" ($C^{i,j}$ are the same as initial C) and the division continues.

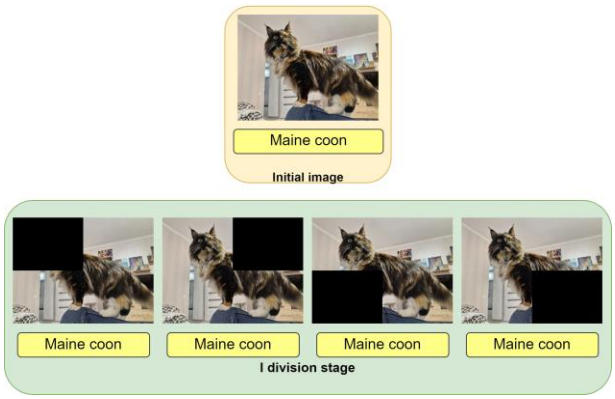

Fig. 2. First division stage

We chose just one of these four images and process only it in our previous RD implementations. Now we switched to breadth first search instead of depth first search of CIP before, and we continue the search until all perturbed images in the current depth level are processed (by default) or other stop condition is met.

Let's look at the first perturbed image and divide it again (let it be $h_m = w_m = 2$), so the size of the hiding rectangle is 4 times less compared to the previous stage. There are 12 such images (Fig. 3) and they are processed one by one again. The third image is classified as Japanese chin, so the complementary image $I - I^{i,j}$ is built and verified. Its classification result is the same as initial $C' = C$ and thus the explanation in the form of complementary images pair is successfully found. If the classification result of complementary image is other than the initial ($C' \neq C$) this pair is not useful anymore and is dropped. Our previous implementations of the method included termination of algorithm at this point and we returned the explanation immediately. For this research we continue to investigate other images searching for multiple CIP (as

it appeared there are three such explanation pairs amongst 12 images).

After that the workflow of the method moves to the second image from Fig. 2, third and fourth. When all perturbed images from the previous stage are processed, the flow may go deeper to the next stage, so the first image from Fig. 3 is used as the initial for further dividing.

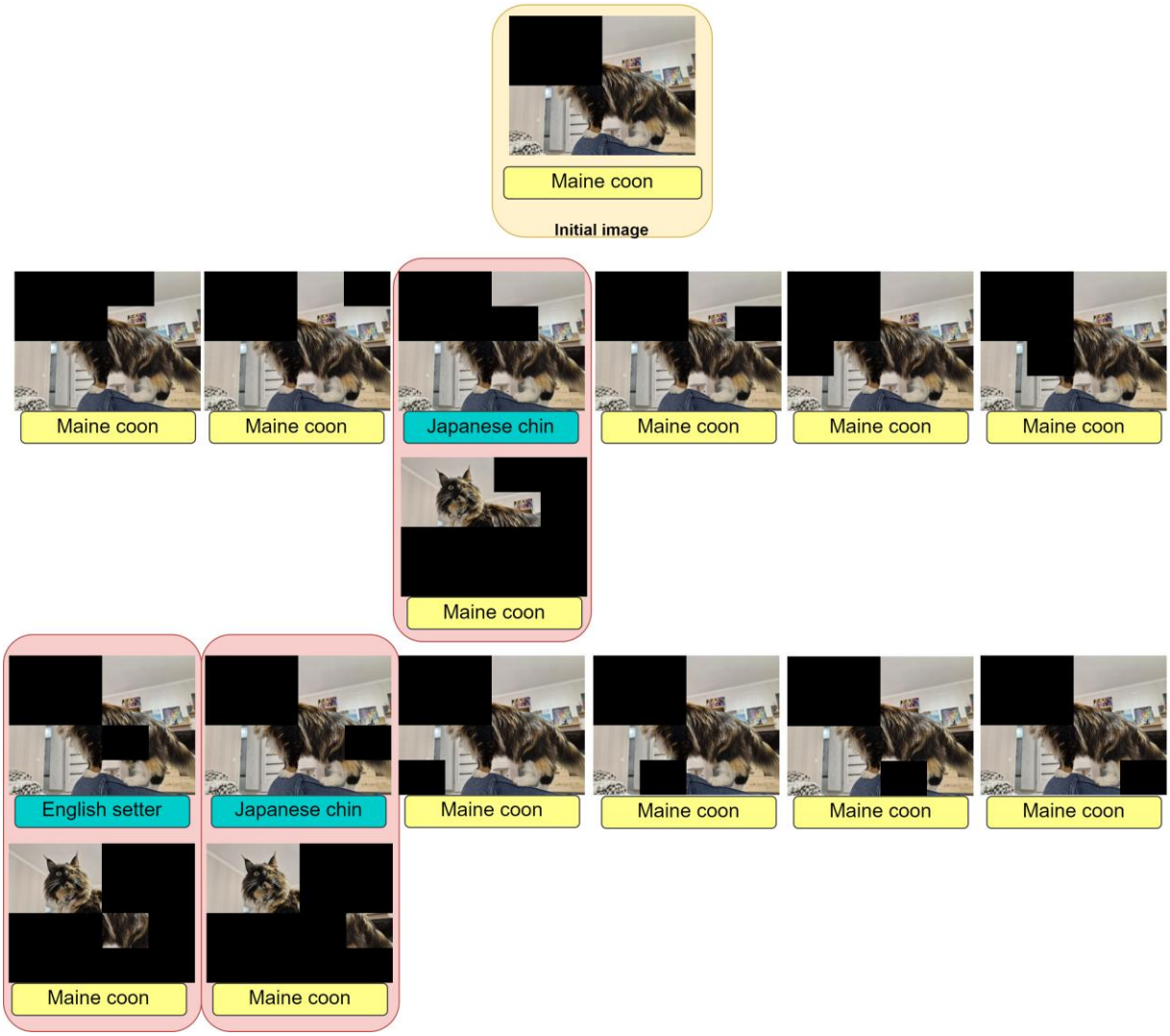

Fig. 3. Second division stage

3.1 Stop criteria

The decision to find multiple CIP requires the processing of a significant number of images, and their quantity grows very quickly, so the criteria to terminate the work of the method should be chosen.

As it comes from [23] the quantity of the perturbed images (assuming again for simplicity that $h_1 = w_1 = 2$ and $h_m = w_m = 2$) at the stage n is

$$2^{2n} - \sum_{k=1}^{n-1} 2^{2k},$$

so, for the example described above, the quantity of images for the each of twelve images perturbed after the stage 2 will be 44 but the size of the hiding rectangle will occupy 64 times less area compared to the size of the original image. Increasing h_1, w_1, h_m, w_m will lead to the even faster growing of this function. There is also important to note that the deeper the stage (n is bigger) – the less our chances are to find CIP, but these CIP are more specific and interesting, capturing small patches of the image.

The overall quantity Q of images for the stage n could be written as a multiplication of quantity of images from previous stage and number of perturbed images P for current stage with the following recursive expressions:

$$Q_n = Q_{n-1}P_{n-1},\ P_n = P_{n-1}h_m w_m - 1,\ n \geq 2,$$

$$Q_1 = h_1 w_1,\ P_0 = h_1 w_1 - 1.$$

Two new decision thresholds were introduced to stop the work of the RD method. First one T_{fail} covers the case when there are still no explanations found, but the quantity of perturbed images is too big for the particular stage. The second threshold T_1 was designed to stop further processing if at least one explanation was

already found. The choice of these thresholds could depend on the performance requirements for the method in the particular case and provide nice flexibility to control the process.

If we set $T_1 = 30$ for the example described above the work of the method stops after the processing of 30 images perturbed taking the first image of Fig. 3 as initial, as there are 44 perturbed images for it in total, and at least one CIP was already found. If we set $T_1 = 45$ that will mean that all 12x44 images for this stage will be processed.

So, the overall description of the RD method requires such hyperparameters to be chosen:

– minimal size of the image fragment to consider as valuable (set to be at least 32 pixels in width and height in all experiments);

– quantity of parts to divide image on each stage – h_1, w_1 for the initial (first) division, h_m, w_m for the all next stages, h_{add}, w_{add} for the additional "last chance" split (if enabled, described in previous papers as a last chance to find the explanation with making one step back and choosing other image from previous stage). We considered only equal integer values to divide the image;

– thresholds to stop processing T_{fail} for the case when there are no explanations found, and T_1 to stop further processing if at least one explanation was already found. We used $T_{fail} = 100$, $T_1 = 30$ as default values. $T_1 = 0$ means to stop processing immediately when first explanation is found (similar to our previous researches).

3.2 Multiple CIP

After RD is completed, we may have multiple CIP each of which could be used as an independent explanation result. A lot of different methods to combine them into one joint explanation could be created, in this section we propose two to find the CIP that has the smallest area.

Let M be the binary mask of the explanation images E found after RD (examples are shown in Fig. 4) that preserves the initial classification result. Let's assume for simplicity that visible pixels from the initial image have value 1 in masks while hidden parts are zeros. So, the area of the mask M of size $n \times m$ could be written as

$$A(M) = \sum_{i=1}^{n} \sum_{j=1}^{m} M^{i,j} .$$

Firstly, we select the single explanation image $E_{\min A}$ with the mask that has the smallest area, if there are few of them, we pick any:

$$E_{\min A} = E_{\arg(\min_i A(M^i))} .$$

It is clear that the area is increased with increasing division stage (RD adds new rectangle when going deeper), so CIP found at the lowest stages will prevail over the more specific ones. It is worth noting that minimization of the area is not a strict requirement, just first brief idea, any image could be selected also in order to find more specific explanation. We expect to find the minimal (not guaranteed formally) explanation region but not the most specific one in this example.

Referring to the picture presented in Fig. 2 and Fig. 3 first four masks have the same area and one of them (the last) is chosen (Fig. 4) for further processing.

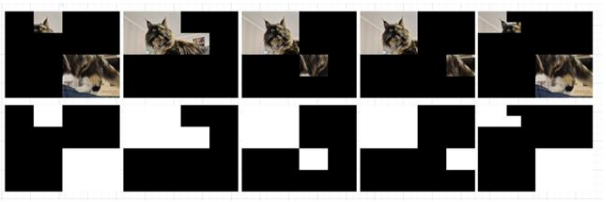

Fig. 4. Explanatory images E (top row) and corresponding masked images M (bottom row) for CIP found after RD

After that we perform SLIC segmentation for the chosen image $E_{\min A}$ searching for approximate 10 segments, and try to remove some superpixels preserving the CIP properties:

$$E_{final} = CIP(SLIC_{10}(E_{\min A})), \quad (1)$$

where $SLIC_{10}()$ means the partition of the masked image into 10 superpixels, and $CIP()$ means the function that checks whether hiding particular superpixels preserves the CIP properties of the explanations.

Let's assume that there were s superpixels found. Firstly, all combinations of $s-1$ superpixels are considered, they are disabled (filled with black) one by one in turn, leaving only one superpixel active, complementary image is created and classification results for new pair is checked.

If they form a CIP – the process is stopped, otherwise all perturbed pairs from $s-2$ superpixels are analyzed and so on. So, actually, this procedure implements one more iterative minimization (again – not strictly formal as sizes of superpixels are not guaranteed to be exactly the same) over the useful mask area with early stopping. This process is illustrated in Fig. 5.

The other way is to apply the abovementioned iterative procedure to all images obtained after RD and minimize the area after SLIC application:

$$E_{final} = \min_i A(CIP(SLIC_{10}(E_i))) . \quad (2)$$

This approach usually needs more time but sometimes allows us to reduce the important area significantly (the chances for this are about 0.35 according to modeling, and in 65% cases both these approaches return same explanation). The example of result is presented in Fig. 6, where the third picture (the first one in the second row) has the minimal area and represents the final CIP.

Searching for multiple explanations made it possible to decrease the quantity of such cases when RD failed to find explanation at all.

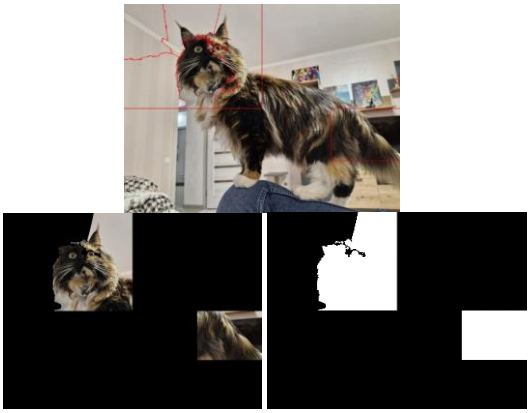

Fig. 5. Reducing the CIP area with SLIC: segmented superpixels (top), reduced part of the image that preserves correct classification (bottom, left), and corresponding mask (bottom, right)

4 Experimental modeling

We used Oxford-IIIT Pet Dataset [25, 26] to train and evaluate different models. It contains 37 cats and dogs breed classes (12 for cats and 25 for dogs), about 200 images per class. This dataset is convenient because each image has an associated ground-truth annotation of breed, region of interest containing head, and background/foreground/undefined pixel masks. The initial training/validation parts of the dataset contain 1846 and 1834 respectively, but we redistribute images to have 3310 of them for the training and 370 (first 10 validation images of each class) for validation purposes. Test set contains 3669 images.

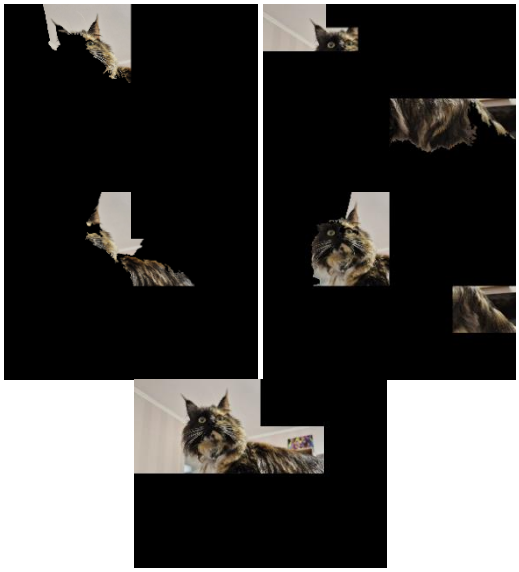

Fig. 6. Image explanations after minimizing mask area over SLIC results

We trained CNN following the transfer learning approach: weights from MobileNet V2 [27] pretrained on ImageNet were used as feature extractor layers followed by dense layers. The architecture of CNN is shown in Fig. 7, it contains the initial input layer for the image scaled to 300x300 pixels, MobileNet V2 convolution layers, global average max pooling layer, and two dense layers. Second dense layer contains 37 neurons (according to quantity of the classes for cats and dogs) with softmax activation function.

Fig. 7. CNN architecture

The training of all models was done using Adam optimizer during 500 epochs (at most). Early stop procedure was implemented to terminate learning after 50 epochs without validation dataset accuracy improving and saving only the best model weights.

The accuracy of the model obtained for the test part of the dataset was 0.8833.

4.1 Search for explanations

In our previous research [21], we used IOU metric and compared the obtained explanation mask with ground-truth masks available for this dataset. But for this paper such comparison seems to be not relevant as the ground-truth masks are of full-size contour for cat/dog breed and we are searching for small regions that preserve classification result. So, we used implementations of Grad-CAM [24] and RISE [14] as baseline models to compare the proposed RD method with.

It is important to note that all experiments were performed for those images only, that were initially classified correctly.

The core of RISE [15] method is the random masking of the initial image, testing classification after that and building the importance map. RISE produces responses of different amplitudes, small for some images, larger for others. The analysis of responses could be a subjective for sure, because for some image that's good to have strong but narrow extremum in the map notifying about specific object, but for another image, probably, the area of the peak should be wider corresponding to the object of other size.

We used such thresholding of the importance map (IM) obtained after RISE:

$$IM_b(i,j) = \begin{cases} 1, \ if \ \ IM(i,j) > \delta \max\limits_{i,j} IM(i,j), \\ 0, \ otherwise, \end{cases}$$

where IM – is the importance map, IM_b – is a binary mask for the importance map, δ – is some threshold defined beforehand.

We tested $\delta = 0.5$, $\delta = 0.7$, and $\delta = 0.8$ thresholds for RISE and found that it is hard to make

decision which one is better, as there are a lot of good and bad binarizations for all cases. Thresholding with $\delta = 0.5$ produces masks with significant importance areas, so we looked at $\delta = 0.7$ or $\delta = 0.8$ cases mainly. Source code [14] with RISE parameters N = 1000, s = 8, $p_1 = 0.5$ was used in experiments. RISE can require a lot of RAM memory to generate 1000 masks for images, so we applied iterative approach here: we tried to generate 1000 masks, if there is not enough memory, tried 800, 600, 500, 400, 300, 200, and 100 masks in a row to get some explanation finally.

We conducted similar experiments for Grad-CAM method and calculated masks over fused heatmap also with thresholds $\delta = 0.7$, and $\delta = 0.8$. The main problem of Grad-CAM application in our experiments was appearance of empty heatmaps (approximately 18% of cases).

The average IOU between masks for Grad-CAM (shortened to GC for formatting), RISE and RD are shown in Table 1. The options set up for RD were $h_1 = w_1 = 2$ and $h_m = w_m = 2$, 10 superpixels for SLIC, $T_{fail} = 100$, $T_1 = 30$. As one can see, changing the threshold leads to sufficient changes of areas even for same method, e.g., RISE after thresholding over 0.7 has level has IOU only 0.48 with RISE thresholded by 0.8. RD methods after optimizing over masks according to (1) and optimizing over SLIC (according to (2)) have the best degree of agreement with IOU 0.77. But commonly Table 1 confirms that different methods produce very different explanation masks.

Table 1 – **Mean IOU between masks**

	GC 0.7	**GC 0.8**	**RISE 0.7**	**RISE 0.8**	**RD w/ (1)**	**RD w/ (2)**
GC 0.7	1	0.43	0.23	0.17	0.19	0.19
GC 0.8		1	0.19	0.16	0.12	0.13
RISE 0.7			1	0.48	0.20	0.21
RISE 0.8				1	0.14	0.15
RD w/ (1)					1	0.77
RD w/ (2)						1

We also looked at few sample experiments where IOU between explanations provided by RD and other methods are close to zero, meaning there are no intersection of explanation areas. The reason of this for Grad-CAM and RD methods is mostly relates to the empty or strange Grad-CAM results. But commonly such cases confirm the interesting and challenging situations of the explanation problems itself.

The examples of non-intersected explanations obtained for the same classification case by different methods are shown in Fig. 8. The input image for the first example is classified as American bulldog. RD method following the procedure (1) with optimizing over masks area found the CIP showing the importance of the bottom right corner for decision making: hiding just it results in classification of German shorthaired breed, and having only that part enabled preserves American bulldog prediction. On the other hand, thresholding the RISE shows that completely another part is important, turning it to black changes the classification result as well (German shorthaired), but complementary image (containing only the top of the dog's head on the black background) is classified as British shorthair breed. So, different parts of the image are important here, and all of them can affect the decision-making process in its own way.

Second and third examples show that the region of the image highlighted with RISE is important as the classification confidence value decreases but not important enough to change the prediction result. The explanations proposed by RD confirm the importance of pixels in the corners of the input image for the decision made by CNN.

Fig. 8. Examples of different explanations having zero IOU (all sample images are from Oxford-IIIT Pet Dataset [25, 26])

4.2 Performance

We have measured the average time required by Grad-CAM, RISE and RD with optimizations (1) and (2) for the single image using first 500 images from test part of the dataset. The results are presented in Table 2. While performance of Grad-CAM and RD seems to be similar in average, it could be very different for same images. Measured times for particular images do not vary a lot for Grad-CAM, but vary from 5 to 200 seconds for different images for RD. At the same time, about 46% of explanations were found within 13 sec. and about 70% were within 25 sec.

Table 2 – **Average seconds per image, sec**

Grad-CAM	**RISE**	**RD with (1)**	**RD with (2)**
25	58	28	32

Conclusions

The paper describes the approach to search for multiple explanations of the particular CNN image classification case. The core of the method is the updated version of recursive division (RD) we used in our previous works. The main idea (except of recursive hiding of rectangles) is to represent explanation as a complementary images pair (CIP) that allows us to visualize the parts of the image which are important enough to change the class of the input image when hidden and at the same time are important enough to preserve the initial classification result when visible. RD does not guarantee that these are the only such pieces though.

RD searches for the explanations with the hiding 1-4 rectangular areas of different sizes, that is complex, and generates a lot of perturbation images during the work. The parameters of RD method are discussed to choose the criteria to stop the processing when few explanations are found or the further processing requires too much time and/or memory resources.

The merging of multiple explanations back to single using SLIC superpixels segmentation applied to the explanations found was proposed. This allowed us to reduce the image explanation area and sometimes find more visually attractive CIP, but such merging is not strictly required if we are satisfied with just multiple CIP explanations.

We compared results of searching for multiple explanation images based on RD with other popular methods like RISE and Grad-CAM. RD appeared to be somewhat better than Grad-CAM (because it finds more successful explanations), and faster than RISE (because RD requires less memory and time on average), but RD has its own drawbacks and the decision on which method is the best depends on a lot of other circumstances.

Acknowledgements

We dedicate this paper and thank the Armed Forced of Ukraine, all related Forces, and divisions, doctors, volunteers, and everyone who supports Ukraine. Low bow and eternal memory to all Defenders.

We thank Olena Valenok for the picture of her Maine coon pet that allowed us to present the idea of our method perfectly.

References

1. Carvalho, D.V., Pereira, E.M., and Cardoso, J.S. (2019), “Machine learning interpretability: A survey on methods and metrics”, *Electronics*, vol. 8(8), 832, doi: https://doi.org/10.3390/electronics8080832
2. Gilpin, L., Bau, D., Yuan, B., Bajwa, A., Specter, M., and Kagal, L. (2018), “Explaining explanations: An overview of interpretability of machine learning”, *2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA)*, pp. 80–89, doi: https://doi.org/10.1109/DSAA.2018.00018
3. Ibrahim, R., and Shafiq, M. (2023), “Explainable convolutional neural networks: A taxonomy, review, and future directions”, ACM Computing Surveys, vol. 55, no. 10, pp. 1–37, doi: https://doi.org/10.1145/3563691
4. Linardatos, P., Papastefanopoulos, V., and Kotsiantis, S. (2021), “Explainable AI: A review of machine learning interpretability methods”, *Entropy*, vol. 23, no. 1, doi: https://doi.org/10.3390/e23010018
5. Molnar C. (2025), *Interpretable Machine Learning*, 3rd rd., available at: https://christophm.github.io/interpretable-ml-book
6. Song, Y. (2020), *Towards multi-scale visual explainability for convolutional neural networks*, available at: http://www.diva-portal.org/smash/get/diva2:1468770/FULLTEXT01.pdf
7. Zhang, Q., Wang, X., Wu, Y., Zhou, H., and Zhu, S. (2021), “Interpretable CNNs for object classification”, *IEEE Transactions on Pattern Analysis & Machine Intelligence*, vol. 43, no. 10, pp. 3416–3431, doi: https://doi.org/10.1109/TPAMI.2020.2982882
8. Zhang, Q., Wu, Y., and Zhu, S. (2018), “Interpretable convolutional neural networks”, *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8827– 8836, doi: https://doi.org/10.1109/CVPR.2018.00920
9. Zhou, J., Gandomi, A.H., Chen, F., and Holzinger, A. (2021), “Evaluating the quality of machine learning explanations: A survey on methods and metrics”, *Electronics*, vol. 10, no. 5, doi: https://doi.org/10.3390/electronics10050593
10. Chalyi, S., Leshchynskyi, V., and Leshchynska, I. (2019), “Designing explanations in the recommender systems based on the principle of a black box”, *Advanced Information Systems*, vol. 3, no. 2, pp. 47–51, doi: https://doi.org/10.20998/2522-9052.2019.2.08
11. Chalyi, S., and Leshchynskyi, V. (2023), “Probabilistic counterfactual causal model for a single input variable in explainability task”, *Advanced Information Systems*, vol. 7, no. 3, pp. 54–59, doi: https://doi.org/10.20998/2522-9052.2023.3.08
12. Lundberg, S., and Lee, S. (2017), “A unified approach to interpreting model predictions”, *The 31st International Conference on Neural Information Processing Systems (NIPS'17)*, pp. 4768–4777, doi: http://doi.org/10.48550/arXiv.1705.07874
13. Ribeiro, M., Singh, S., and Guestrin, C. (2016), “Why should I trust you? Explaining the predictions of any classifier”, *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining ACM 2016*, pp. 1135–1144, doi: https://doi.org/10.1145/2939672.2939778
14. (2025), *Randomized Image Sampling for Explanations (RISE)*, available at: https://github.com/eclique/RISE/blob/master/Easy_start.ipynb
15. Petsiuk, V., Das, A., and Saenko, K. (2018), “RISE: Randomized input sampling for explanation of black-box models”, *ArXiv*, doi: https://doi.org/10.48550/arXiv.1806.07421
16. Selvaraju, R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. (2017), “Grad-CAM: Visual explanations from deep networks via gradient-based localization”, *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 618–626, doi: https://doi.org/10.1109/ICCV.2017.74
17. Jiang, M., Khorram, S., and Fuxin, L. (2024), “Comparing the Decision-Making Mechanisms by Transformers and CNNs via Explanation Methods”, *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9546–9555, doi: https://doi.org/10.1109/CVPR52733.2024.00912
18. Shitole, V., Fuxin, L., Kahng, M., Tadepalli, P., and Fern, A. (2021), “One explanation is not enough: structured attention

graphs for image classification", *Proceedings of the 35th International Conference on Neural Information Processing Systems (NIPS '21)*, vol. 34, pp. 11352–11363, available at: https://dl.acm.org/doi/10.5555/3540261.3541129
19. Gorokhovatskyi, O., and Peredrii, O. (2020), "Multiclass image classification explanation with the complement perturbation images", *Data Stream Mining & Processing (DSMP 2020), Communications in Computer and Information Science,* vol. 1158, pp. 275–287, doi: https://doi.org/10.1007/978-3-030-61656-4_18
20. Gorokhovatskyi, O., and Peredrii, O. (2021), "Recursive division of image for explanation of shallow CNN models", *Pattern Recognition. ICPR International Workshops and Challenges Proceedings*, Part III, pp. 274–286, doi: https://doi.org/10.1007/978-3-030-68796-0_20
21. Gorokhovatskyi, O., and Peredrii, O. (2024), "Recursive Division Explainability as a Factor of CNN Quality", *Lecture Notes on Data Engineering and Communications Technologies*, vol. 219, pp. 308–325, doi: https://doi.org/10.1007/978-3-031-70959-3_16
22. Gorokhovatskyi, V., Chmutov, Y., Tvoroshenko, I., and Kobylin, O. (2025), "Reducing Computational costs ny compressing the structural description in image classification methods", *Advanced Information Systems*, vol. 9, no. 1, pp. 5–12, doi: https://doi.org/10.20998/2522-9052.2025.1.01
23. Gorokhovatskyi, O., Peredrii, O., and Gorokhovatskyi, V. (2020), "Interpretability of Neural Network Binary Classification with Part Analysis", *The Third IEEE International Conference on DataStream Mining & Processing*, pp. 136–141, doi: https://doi.org/10.1109/DSMP47368.2020.9204310
24. (2025), *Grad-CAM - training tutorial*, available at: https://colab.research.google.com/drive/1rxmXus_nrGEhxlQK_By38AjwDxwmLn9S?usp=sharing
25. Parkhi, O., Vedaldi, A., Zisserman, A. and Jawahar, C. V. (2025), *The Oxford-IIIT Pet Dataset*, available at: https://www.robots.ox.ac.uk/~vgg/data/pets/
26. Parkhi, O., Vedaldi, A., Zisserman, A., and Jawahar, C. (2012), "Cats and dogs", *2012 IEEE Conference on computer vision and pattern recognition (CVPR)*, pp. 3498–3505, doi: https://doi.org/10.1109/CVPR.2012.6248092
27. Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., and Chen, L. (2018), "MobileNetV2: Inverted residuals and linear bottlenecks", *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4510–4520, doi: https://doi.org/10.1109/CVPR.2018.00474

Received (Надійшла) 11.03.2025
Accepted for publication (Прийнята до друку) 18.06.2025

ABOUT THE AUTHORS / ВІДОМОСТІ ПРО АВТОРІВ

Гороховатський Олексій Володимирович – кандидат технічних наук, доцент кафедри інформатики та комп'ютерної техніки, Харківський національний економічний університет імені Семена Кузнеця, Харків, Україна;
Oleksii Gorokhovatskyi – PhD, Associate Professor, Department of Informatics and Computer Engineering, Simon Kuznets Kharkiv National University of Economics, Kharkiv, Ukraine;
e-mail: oleksii.gorokhovatskyi@gmail.com; ORCID Author ID: https://orcid.org/0000-0003-3477-2132;
Scopus ID: https://www.scopus.com/authid/detail.uri?authorId=23099879900.

Передрій Олена Олегівна – кандидат технічних наук, доцент кафедри інформатики та комп'ютерної техніки, Харківський національний економічний університет імені Семена Кузнеця, Харків, Україна;
Olena Peredrii – PhD, Associate Professor, Department of Informatics and Computer Engineering, Simon Kuznets Kharkiv National University of Economics, Kharkiv, Ukraine;
e-mail: olena.peredrii@hneu.net; ORCID Author ID: https://orcid.org/0000-0003-0390-1931;
Scopus ID: https://www.scopus.com/authid/detail.uri?authorId=57202751577.

Тесленко Олег Володимирович – кандидат технічних наук, доцент кафедри інформатики та комп'ютерної техніки, Харківський національний економічний університет імені Семена Кузнеця, Харків, Україна;
Oleh Teslenko – PhD, Associate Professor, Department of Informatics and Computer Engineering, Simon Kuznets Kharkiv National University of Economics, Kharkiv, Ukraine;
e-mail: oleh.teslenko@hneu.net; ORCID Author ID: https://orcid.org/0000-0003-3105-9323;
Scopus ID: https://www.scopus.com/authid/detail.uri?authorId=57210286707.

Множинні пояснення рекурсивним поділом для проблем класифікації зображень

О. В. Гороховатський, О. О. Передрій, О. В. Тесленко

Анотація. Мета дослідження. У цій статті пропонується підхід до пошуку множинних пояснень випадку класифікації зображень CNN. **Результати дослідження.** Основою методу є рекурсивний поділ (RD), який виконує збурення вхідного зображення з приховуванням різних прямокутних частин. Пояснення представлено у вигляді пари зображень, які доповнюють один одного (CIP): така пара зображень дозволяє нам візуалізувати частини зображення, які є достатньо важливими, щоб змінити клас вхідного зображення, коли вони приховані, і водночас є достатньо важливими, щоб зберегти початковий результат класифікації, коли вони видимі. Обговорюються параметри методу RD для вибору критеріїв зупинки обробки, коли знайдено декілька пояснень або подальша обробка вимагає забагато часу та/або ресурсів пам'яті. Було запропоновано два підходи до об'єднання кількох CIP назад в одне пояснення за допомогою сегментації SLIC. Вони дозволили нам зменшити корисну область пояснення зображення та іноді знайти візуально привабливіші CIP порівняно з попередніми реалізаціями RD. Таке об'єднання не є суворо обов'язковим, якщо достатньо лише кількох пояснень CIP для аналізу CNN. **Висновки.** Реалізацію запропонованого підходу для задачі класифікації порід котів та собак було порівняно з іншими популярними методами, такими як RISE та Grad-CAM, обговорено переваги та недоліки. Аналіз ефективності підтвердив перевагу запропонованих методів, оскільки вони порівнянні або швидші за відомі та дозволяють знаходити зображення з кількома поясненнями.

Ключові слова: класифікація зображення; CNN; AI; «чорна скриня»; пояснювальність; збурення зображення; приховування частин; рекурсивний поділ; пара доповнюючих зображень (CIP).