Mitigating Bias: Enhancing Image Classification by Improving Model Explanations

Raha Ahmadi* RAHAAHMADI@AUT.AC.IR
Mohammad Javad Rajabi* RAJABI2001@AUT.AC.IR
Mohammad Khalooei KHALOOEI@AUT.AC.IR
Amirkabir University of Technology
Mohammad Sabokrou MOHAMMAD.SABOKROU@OIST.JP
Okinawa Institute of Science and Technology
Institute for Research in Fundamental Sciences(IPM)

Editors: Berrin Yanıkoğlu and Wray Buntine

Abstract

Deep learning models have demonstrated remarkable capabilities in learning complex patterns and concepts from training data. However, recent findings indicate that these models tend to rely heavily on simple and easily discernible features present in the background of images, rather than the main concepts or objects they are intended to classify. This phenomenon poses a challenge to image classifiers as the crucial elements of interest in images may be overshadowed. In this paper, we propose a novel approach to address this issue and improve the learning of main concepts by image classifiers. Our central idea revolves around concurrently guiding the model's attention toward the foreground during the classification task. By emphasizing the foreground, which encapsulates the primary objects of interest, we aim to shift the focus of the model away from the dominant influence of the background. To accomplish this, we introduce a mechanism that encourages the model to allocate sufficient attention to the foreground. We investigate various strategies, including modifying the loss function or incorporating additional architectural components, to enable the classifier to effectively capture the primary concept within an image. Additionally, we explore the impact of different foreground attention mechanisms on model performance and provide insights into their effectiveness. Through extensive experimentation on benchmark datasets, we demonstrate the efficacy of our proposed approach in improving the classification accuracy of image classifiers. Our findings highlight the importance of foreground attention in enhancing model understanding and representation of the main concepts within images. The results of this study contribute to advancing the field of image classification and provide valuable insights for developing more robust and accurate deep-learning models.

Keywords: Deep learning; image classification; foreground attention; concept learning; model enhancement.

1. Introduction

Deep neural networks (DNNs) have gained widespread adoption in various computer vision tasks due to their superior performance and remarkable capabilities in learning complex

*. Equal contribution

© 2023 R. Ahmadi*, M.J. Rajabi*, M. Khalooei & M. Sabokrou.

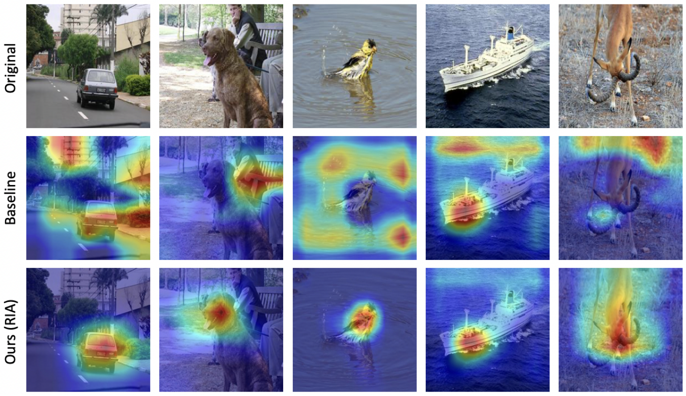

Figure 1: Our approach demonstrates enhanced robustness of the Grad-CAM explanation and reduced attention to background regions, resulting in more accurate and focused explanations.

patterns and concepts from training data. Nevertheless, recent studies have revealed that DNNs generally learn the most uncomplicated features, like biases, and do not concentrate well on the main key objects of input image Mehmanchi et al. (2023); Singh et al. (2020). Additionally, most of the computer vision classification tasks experienced similar issues, where the models were biased towards the non-central objects or the background of the primary object in the input sample. Moayeri et al. (2022) has discovered that highly precise models can be greatly affected by background noise, more so than foreground noise. This indicates that in complex scenarios, deep learning models unexpectedly depend on non-central objects or features in the background rather than accurately identifying the primary key objects that were programmed to be classified. This poses a significant challenge to the reliability of DNNs.

In general, the interpretability of DNNs has thus emerged as an important research area, with the purpose of enhancing transparency and applicability of DNNs. This makes it possible to provide trustfully and helps identify any spurious correlations the network may have inadvertently learned to use to make its decision Singh et al. (2020). The use of Explainable Artificial Intelligence (XAI) approaches can be considered valuable for interpreting DNNs and pointing out biases. It is unfortunate that methods like Grad-CAM Selvaraju et al. (2017), have proven to be unreliable. For instance, when a cat image and its rotated version are given to the classifier; it recognizes them as a cat, but their explainability results from Grad-CAM differ. Pillai et al. (2022) raises this concern and proposes a training method inspired by contrastive self-supervised learning to address this issue.

To address challenges and prevent bias toward the background in image classification, We propose aligning the Grad-CAM interpretations mechanism with the main objects of interest. Inspired by self-supervised learning, our method uses a novel loss function to align the attention area of the model with the main object. This approach aims to enhance trustworthiness, reduce sensitivity to background noise, and improve the model's interpretability in complex computer vision tasks.

We evaluated our proposed Region of Interest Activation (RIA) method to assess its impact on classification accuracy and the sensitivity of the model to the foreground and background attributes. Our method aims to guide the model in prioritizing the most discriminative features of the input image, thereby enhancing accuracy and reliability in fine-grained classification scenarios.

Figure 1 illustrates some examples which demonstrate that the RIA method significantly improves the interpretation of the model in primary objects of corresponding input images, as identified by an object detector. Furthermore, we observed that our method enhances classification accuracy in fine-grained settings and when faced with foreground and background noise. The evaluation highlights the effectiveness of our RIA method in improving the accuracy and reliability of deep learning models for image classification tasks. By directing models to focus on relevant features and reducing sensitivity to irrelevant attributes, our method has the potential to enhance performance across various applications reliant on image classification.

Our main contributions are summarized as follows:

1. Our study highlights the significance of how models can be biased towards objects or features in the background, which can impact the model's explanations and lead to inaccurate decisions.

2. We propose a novel approach that promotes the model's focus on the primary object of the image by utilizing the proximity between the model's attention area and the detected area of an object.

3. We have adapted the basic IoU loss to account for the anticipated inaccuracies in estimating the area of an object by an object detector.

4. Our approach not only enhances the accuracy of the model but also significantly improves its reliability and additionally would be more robust under the foreground and background noises.

2. Related Works

2.1. Interpretability Methods

Interpretability methods for deep neural networks have been extensively studied in recent years, driven by the widespread adoption of these models across various tasks. Ribeiro et al. (2016) have highlighted that machine learning models often capture undesirable correlation artifacts during training, which can be challenging to identify solely by relying on prediction accuracy. In order to address this issue, several methods have been proposed to detect salient regions in images. Zeiler and Fergus (2014) introduced an approach that leveraged gradients of the class conditional output with respect to the input image. By identifying spatial locations with large gradient magnitudes, a saliency map corresponding to the class could be obtained. Building upon this work, Springenberg et al. (2014) and Sundararajan et al. (2017) further enhanced the quality of saliency maps, resulting in sharper visualizations. Ross et al. (2017) showed that constraining the gradient explanations to be small in irrelevant areas using an annotation mask improved the quality of these explanations, albeit with additional computational costs. Class Activation Mapping (CAM) was introduced by Zhou et al. (2016) to produce a coarse localization heatmap by utilizing a global average pooling (GAP) layer to calculate the gradients flowing into the final convolution layer. Gradient-weighted Class Activation Mapping (Grad-CAM), proposed by Selvaraju et al. (2017), extends the concept of CAM by utilizing gradients flowing into the final convolutional layer for a specific class. This approach generates a coarse localization map that highlights the important regions in the image for predicting the corresponding class.

Our proposed method is also based on Grad-CAM heatmaps. However, we introduce a novel approach by using Grad-CAM during the training process. This allows us to guide the generation of heatmaps, resulting in improved explanations that align more accurately with the known regions describing the desired features.

2.2. Explanation-guided learning

In recent studies, researchers have explored the integration of explanations during model training to enhance predictive performance. Rieger et al. (2020) aims to align explanations with human annotations based on domain knowledge. This strategy helps reduce the model's reliance on background pixels. However, obtaining ground truth explanations can be labor-intensive Wang et al. (2020) or even unfeasible due to subjectivity in real-world tasks Roscher et al. (2020). Pham et al. (2021) employed segmentation masks to direct attention maps towards important regions of images for attribute prediction tasks. Selvaraju et al. (2021) utilized saliency maps generated with DeepUSPS Nguyen et al. (2019) during training to guide attention maps, aiming to enhance self-supervised representation learning. Similarly, Pillai et al. (2022) focused on maintaining consistent explanations to facilitate generic representation learning through the use of contrastive objectives. The goal of such approaches is not only to improve performance but also to make sure that the model is "right for the right reasons" Ross et al. (2017). For classifiers, this typically involves jointly optimizing both classification performance and localization to object features.

In this work, we present a novel loss function that guides DNNs toward accurate object localization and enhances their robustness by reducing reliance on spurious features and background information.

3. Proposed Method

The main goal of our approach is to ensure fairness and reduce biases in image classification while maintaining the consistent interpretability of the model. To achieve this, we propose the Region of Interest Activation Loss (RIA), which encourages the model to classify images accurately and focus on the main concept or object within the images. Our proposed method consists of two key components: (1) Categorical Cross-Entropy Loss (L_{CE}), and (2) RIA: Region of Interest Activation Loss (RIA). These components work together to enhance the network's ability to attend to both foreground and background objects. To begin the learning process, we start with the L_{CE} loss function, which is a standard approach for classification tasks. This loss function guides the network to minimize the discrepancy between predicted and ground truth labels. However, to ensure that the model attends to the foreground objects as well as the background, we introduce the RIA loss function. This loss function leverages the concept of Grad-CAM (Gradient-weighted Class Activation

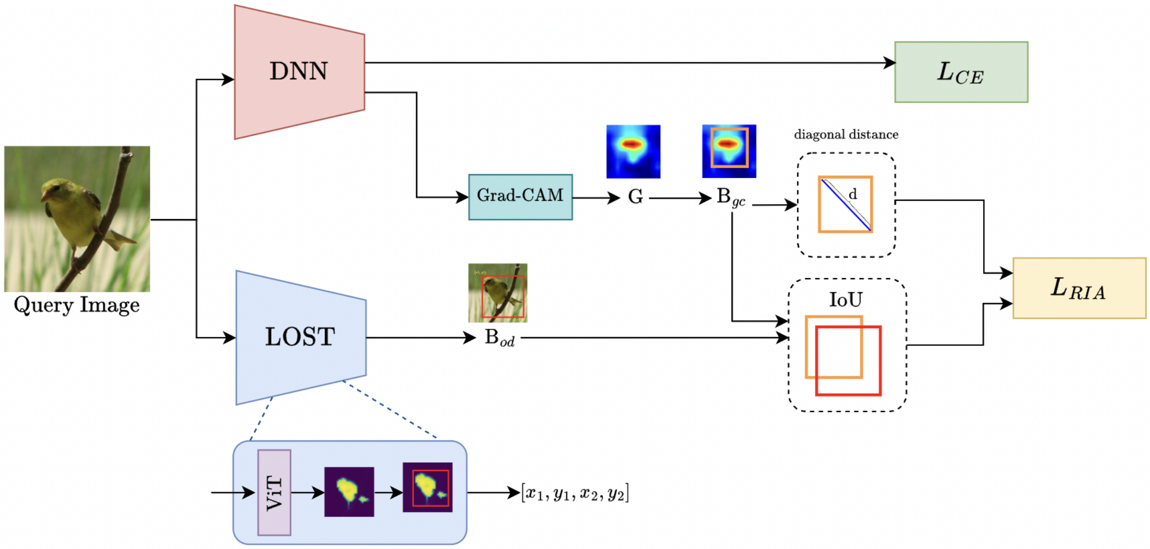

Figure 2: **The block diagram of our method.** Our method consists of both cross-entropy loss L_{CE} and Region of Interest Activation loss L_{RIA} which represents a modified IoU loss. We load a batch of images and consider one to be the query image. We feed the query image to the network and calculate L_{CE}. We calculate Grad-CAM for this image on the top predicted category and obtain a bounding box from the Grad-CAM heatmap. We feed the query image to the object detector network and get a predicted bounding box. We then compute the loss between these two boxes L_{RIA} and combine it with L_{CE} loss.

Mapping) and aims to maximize the intersection over the union of the foreground object and the Grad-CAM during training. By incorporating RIA, we encourage the network to focus on relevant regions of interest and improve its ability to distinguish foreground and background objects. To combine these losses effectively, we define a new objective function and learning scenario. This combined objective function encompasses both the L_{CE} loss and the RIA loss, allowing the model to optimize for classification accuracy while also attending to important regions in the input data. The learning scenario incorporates these losses throughout the training process, enabling the model to learn the necessary representations. It is worth mentioning that, we exploited an unsupervised and low-cost object detector.

Figure 2 illustrates the block diagram of our method. This section provides a brief overview of the Grad-CAM interpretation algorithm and the object detection algorithm. We then delve into the details of the Region of Interest Activation Loss term.

3.1. Background on Grad-CAM

Grad-CAM uses the gradient information flowing into the last convolutional layer of the CNN to assign importance values to each neuron for a particular decision of interest. Selvaraju et al. (2017) To apply Grad-CAM to an input image x and a deep neural network f, we start by obtaining the output logits y for each category by feeding x to the model, where y^t corresponds to the output for category t, and feeding y through a SoftMax operator produces the probability distribution over categories. We then select last convolutional layer of the network and compute the derivative of the predicted output with respect to each channel of the convolutional layer, averaged over all spatial locations to get the importance of each channel of the convolutional layer in making the current prediction:

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

where A_{ij}^k is the activations of the convolutional layer at channel k and location (i, j), Z is a normalizer, and α_k^t is the importance weights of channel k. Then, we perform a weighted combination of forward activation maps for each channel to get a 2D matrix over spatial locations, and follow it by a ReLU to discard negative values.

$$gradcam_{ij}^t = Relu(\sum k \alpha_k^t A_{ij}^k) \tag{2}$$

Finally, we resize it using bilinear interpolation to the size of the input image to get the interpretation heatmap.

3.2. Obtaining Bounding Box from heatmap

To obtain a bounding box (B_{gc}) from the heatmap generated by Grad-CAM (G), we apply threshold segmentation (T) on G as output of Grad-CAM to create Grad-CAM binary mask (GBM), where regions of interest are identified based on intensity values above the threshold.

$$GBM(x, y) = \begin{cases} 1, & \text{if G}(x, y) > \text{T} \\ 0, & \text{otherwise} \end{cases} \tag{3}$$

Then, we have employed connected component analysis to identify individual connected regions within the binary mask. Each connected region corresponds to a potential object. For each connected component, we calculate the bounding box by determining the minimum and maximum coordinates in both the horizontal and vertical directions. These coordinates specify the rectangular region that tightly encapsulates the object of interest. By following these steps, we can effectively extract the bounding box associated with the highest score in the heatmap. This bounding box allows us to precisely locate and highlight the region in the input image that played a crucial role in the neural network's classification decision.

3.3. Unsupervised Object Detector

To obtain the target bounding boxes, we use LOST (i.e., Localizing Objects with Self-Supervised Transformers) Siméoni et al. (2021) which can unsupervisedly detect the objects. The LOST method utilizes self-supervised learning techniques, specifically self-supervised transformers, to train a model to discover and localize objects within an image. By localizing objects in image collections without supervision, we can improve the model's ability to become object-agnostic.

In the LOST approach, the image is divided into equal-sized patches and fed into a transformer model. Instead of focusing on the CLS token (i.e., An additional, learned vector called the class token), the key component of the last attention layer is used for computing similarities between different patches. By doing so, we can localize a part of an object by selecting the patch with the least number of similar patches, which is referred to as the seed. The justification for this seed selection criterion is based on the empirical observation that patches of foreground objects are less correlated than patches corresponding to the background. Then the seed is expanded by adding other patches that are highly correlated to it and are thus likely to be part of the same object, a process which is called seed expansion. Finally, a binary object segmentation mask is constructed by computing the similarities of each image patch to the selected seed patches. Then the bounding box of an object is considered as the box that tightly encloses the largest connected component in this mask that contains the initial seed.

3.4. RIA: Region of Interest Activation Loss

To enhance the model's consistency and interpretability, we propose a modified Intersection over Union (IoU) loss that bridges the gap between the object detector's bounding box predictions and the bounding boxes obtained from Grad-CAM.

Generally, the IoU-based loss can be defined as

$$IoU_{Loss} = 1 - IoU + R(B, B_{gt}) \tag{4}$$

where R(B, B_{gt}) is the penalty term for predicted box B and target box B_{gt} Zheng et al. (2020).

IOU is a commonly used evaluation metric in computer vision and object detection tasks. It is used to measure the similarity between two arbitrary shapes (boxes). Generally, the IoU metric is defined as

$$IoU = \frac{|B \cap B_{gt}|}{|B \cup B_{gt}|} \tag{5}$$

In our approach, we use the bounding box attained from Grad-CAM as the predicted box(B_{gc}) and the bounding box generated by the object detector as the target box(B_{od}). Our goal is to minimize the differences between the predicted and target boxes, making them as similar as possible. However, the generated object detector boxes as the target boxes may not be entirely accurate and flawless which is not desired. These bounding boxes may be larger than necessary, covering not only the main object but also including some of the background objects. Therefore, it is important to acknowledge that these bounding boxes are not entirely error-free and may require further refinement. To address this issue, we are considering a modification to the IoU term. Specifically, our IoU use only the box B_{gc} in the denominator instead of the union of B_{gc} and B_{od}.

$$I\hat{o}U = \frac{|B_{od} \cap B_{gc}|}{|B_{gc}|} \tag{6}$$

This term is designed to increase the size of the intersection area without increasing the size of B_{gc}, thereby making the predicted bounding box equal to or smaller than the target bounding box. Additionally, this term promotes the containment of B_{gc} within B_{od}. As a result, the numerator and denominator of $I\hat{o}U$ are identical when the intersection between the predicted(B_{gc}) and target bounding boxes(B_{od}) is equal to B_{gc} itself. We have found that this modification adequately compensates for any inaccuracy of the object detector, since the intersection area between the B_{gc} and B_{od} usually covers the main object. Our model seeks to ensure that the intersection area captures the most significant part of the main object, while excluding any irrelevant objects or background areas.

The size of the Grad-CAM box plays a crucial role in model performance. Large attention regions may encompass unnecessary features, leading to inaccurate predictions. It is important to encourage the model to focus on the most relevant regions by constraining the size of Grad-CAM boxes. To address the issue of large attention regions and improve model performance, we propose a penalty-based modification to the loss function. This modification encourages the model to minimize the size of its attention region, resulting in smaller Grad-CAM boxes that focus on the most discriminative features.

We introduce a diagonal distance penalty for the Grad-CAM box which refers to the length of the diagonal line that spans across the box. It is calculated by measuring the Euclidean distance between the opposite corners of the box. The diagonal distance provides an estimate of the size and extent of the Grad-CAM box in terms of its spatial coverage. By this penalty, our objective is to minimize the size of the predicted bounding boxes while ensuring that they still cover the most significant part of the main object. By doing so, we aim to mitigate the influence of irrelevant objects or background areas on the model's explainability and accuracy.

$$R(B_{gc}) = \lambda * diagonal\ distance(B_{gc}) \tag{7}$$

where λ is the regularization coefficient. With the modifications to the IoU loss and the inclusion of the diagonal distance penalty term, we have now finalized our Interest Activation loss (L_{RIA}) as:

$$L_{RIA} = 1 - I\hat{o}U + R(B_{gc}) \tag{8}$$

Our final loss is the combination of the standard cross-entropy loss (L_{CE}) and our Region of Interest Activation loss (L_{RIA}). Hence we minimize the following loss function:

$$L = \alpha * L_{CE} + \beta * L_{RIA} \tag{9}$$

where α and β are hyper-parameters that control the trade-off between the two loss terms.

4. Experiment Results

In this section, we present the results of a series of experiments conducted using our proposed method on variant models. The main objective of these experiments is to evaluate the impact of our Region of Interest Activation (RIA) loss on the overall performance and robustness of our models, particularly under the addition of background and foreground noise. For each experiment, we conducted a comparative analysis between two models: the baseline model, trained solely with the standard cross-entropy loss, and the consistent model, trained with both the standard cross-entropy loss and the RIA loss. By comparing the performance of these two models, as shown in Table 1 and 2, we can effectively assess the effectiveness of our proposed method in improving classification accuracy and model trustability.

4.1. Dataset

We use RIVAL10 dataset Moayeri et al. (2022), whose samples include RIch Visual Attributions with Localization. RIVAL10 consists of images from 20 categories of ImageNet-1k Deng et al. (2009), with a total of 26k high resolution images organized into 10 classes, matching those of CIFAR10.

4.2. Implementation Details

We use PyTorch Paszke et al. (2019) to train and evaluate our models for all experiments. We use pretrained Resnet18 and Resnet50 models He et al. (2016) with the settings used in a previous publication Moayeri et al. (2022). Additionally, we trained VGG16 Simonyan and Zisserman (2014), Resnet18, and Resnet50 models from scratch. To train our models from scratch on the RIVAL10 dataset, we use a two-step training procedure. Initially, we train the models for 10 epochs without the RIA loss, allowing them to gain preliminary insights into the images. Subsequently, we incorporate the RIA loss and continue training to further refine the models using the guidance provided by the loss function. For training our models from scratch on the RIVAL10 dataset, we use Adam optimizer with a learning rate of 0.001 for Resnets and SGD with a learning rate of 0.01 for VGG16. By hyperparameter tuning, we set $\alpha = 1$, $\beta = 0.5$, $\lambda = 0.1$, and $T = 0.5$ for all experiments using our method. Models were trained on an Nvidia RTX 3090 GPU over 50 epochs.

Model	**Baseline Acc (%)**	**Ours (RIA) Acc (%)**
VGG16	86.6%	**91.7%**
ResNet50	88.46%	**88.99%**
ResNet18	88.51%	**88.78%**

Table 1: Classification Accuracy for models trained from scratch on RIVAL10 validation set.

Model	**Baseline Acc (%)**	**Ours (RIA) Acc (%)**
ResNet50	88.46%	**88.99%**
ResNet18	98.8%	**99%**

Table 2: Classification Accuracy for pretrained models on RIVAL10 validation set.

4.3. Evaluating the Sensitivity to Background/Foreground

To assess the robustness of our method, we conducted a thorough analysis by adding Gaussian noise to both the foreground and background regions separately. This allowed us to evaluate how the corruption of each region affects the performance of our models. The evaluation of sensitivity to background and foreground in both pretrained models and models trained from scratch, as shown in Figure 3 and 4, indicates that our method outperforms the baseline approach in noisy conditions. The use of RIA loss enables us to mitigate biases towards the background and enhance foreground attention. Consequently, our models exhibit a smaller decrease in classification accuracy compared to the baseline models.

To quantify the sensitivity of a model to foregrounds relative to its sensitivity to backgrounds, we introduce relative foreground sensitivity (RFS). Moayeri et al. (2022) Let a_{fg} and a_{bg} denote accuracy under noise in the foreground and background, respectively. We then define RFS as

$$RFS = a_{bg} - a_{fg} \tag{10}$$

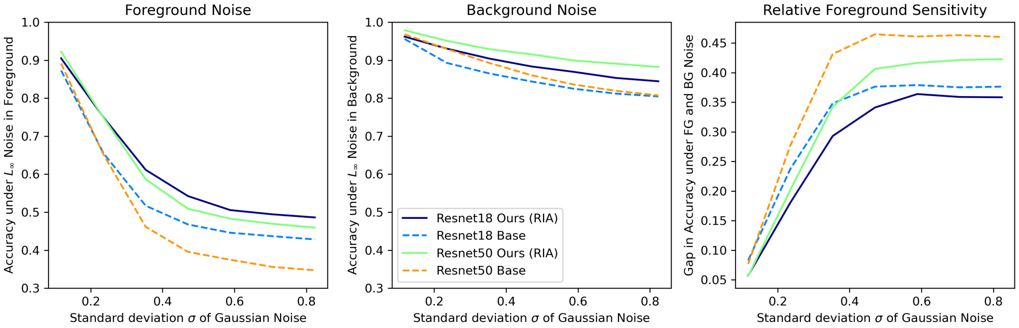

Figure 3: The chart on the left displays accuracy levels under foreground (left) and background (middle) noise at different levels. The models are categorized by their architecture and training method for pretrained models, and each curve represents the average accuracy of all models in the group. On the right, the chart shows the RFS by group.

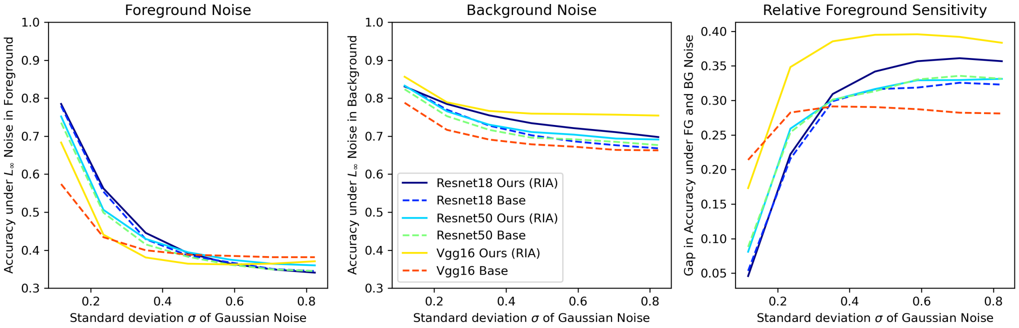

Figure 4: The chart on the left displays accuracy levels under foreground (left) and background (middle) noise at different levels. The models are categorized by their architecture and training for models trained from scratch, and each curve represents the average accuracy of all models in the group. On the right, the chart shows the RFS by group.

4.4. Model Explanation

Our experiments, as shown in Figure 5, 6, 7, have shown that our proposed approach of encouraging the model to focus on the primary object leads to a significant improvement

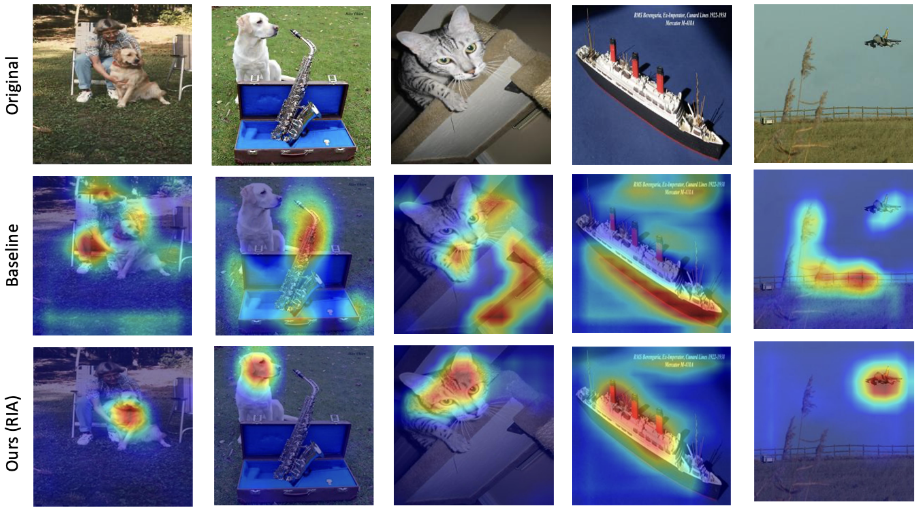

Figure 5: Grad-CAM visualization results for images from RIVAL10 validation set using VGG16 model trained from scratch

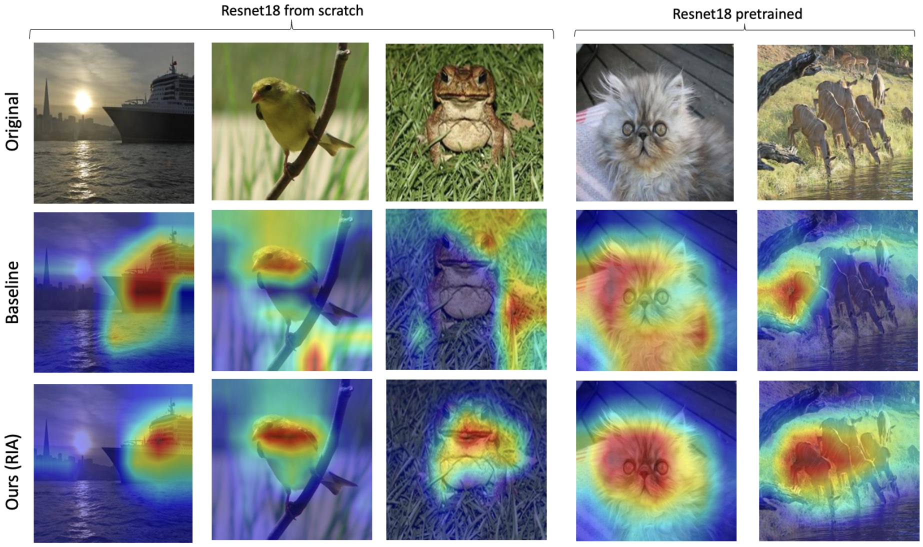

Figure 6: Grad-CAM visualization results for images from RIVAL10 validation set using Resnet18

in the interpretability and clarity of the model's decision-making process. By prioritizing the primary object, the model can extract more relevant features and make informed decisions, which is crucial in various computer vision applications. Moreover, our method has demonstrated the ability to reduce bias towards subsidiary objects and concentrate on the desired object even when it is not discernible enough for model. As a result, our approach eliminates bias towards environmental factors, allowing the model to make accurate deci-

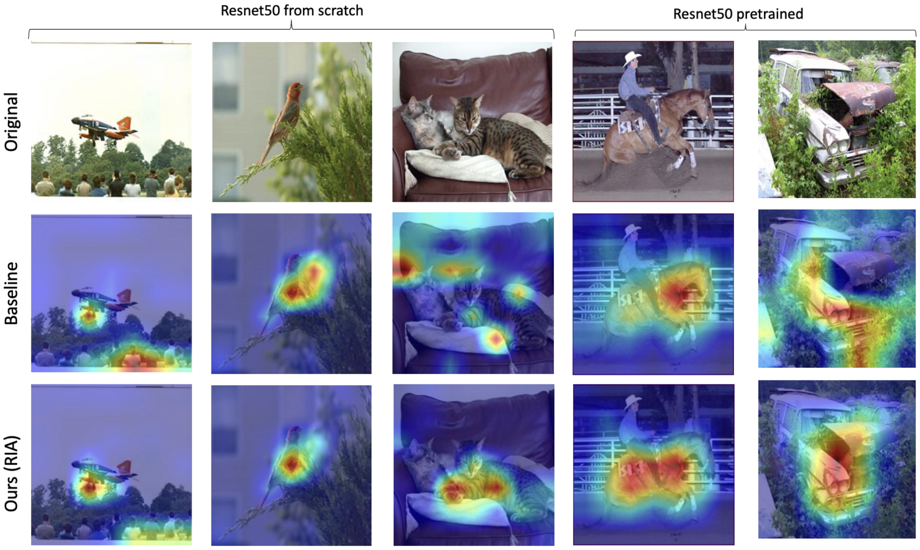

Figure 7: Grad-CAM visualization results for images from RIVAL10 validation set using Resnet50

sions regardless of the object's location and the presence of other objects. For instance, the model is no longer influenced by water when the object is in water, or by branches when identifying birds, resulting in more robust and reliable explanations.

This improvement is significant as it enhances the reliability and robustness of the model, making it more suitable for real-world scenarios where the reliability of the model is critical, and a wrong decision can have severe consequences. Additionally, the model is trustworthy and no longer affected by environmental factors, which ensures that it can make accurate decisions regardless of the object's location or the presence of other objects.

5. Conclusion

We propose Rigion of Interest Activation (RIA), a novel learning approach that improves the interpretability of deep neural networks by encouraging the model to focus on the primary object's area as much as feasible. We emphasize the importance of evaluating the network based on its quality of explanation, and not only classification accuracy. Our RIA method significantly improves the explanation heatmaps while achieving comparable classification accuracy on RIVAL10 dataset. Additionally, our method can enhance the robustness under foreground and background noises while improving the explanation heatmaps and making the model trustworthy. This demonstrates that our method acts as a regularizer that focuses more attention on the discriminating aspects of the image.

References

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and* *pattern recognition*, pages 248–255. Ieee, 2009.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

Mohammad Mahdi Mehmanchi, Mahbod Nouri, and Mohammad Sabokrou. Revealing model biases: Assessing deep neural networks via recovered sample analysis, 2023.

Mazda Moayeri, Phillip Pope, Yogesh Balaji, and Soheil Feizi. A comprehensive study of image classification model sensitivity to foregrounds, backgrounds, and visual attributes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19087–19097, 2022.

Tam Nguyen, Maximilian Dax, Chaithanya Kumar Mummadi, Nhung Ngo, Thi Hoai Phuong Nguyen, Zhongyu Lou, and Thomas Brox. Deepusps: Deep robust unsupervised saliency prediction via self-supervision. *Advances in Neural Information Processing Systems*, 32, 2019.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

Khoi Pham, Kushal Kafle, Zhe Lin, Zhihong Ding, Scott Cohen, Quan Tran, and Abhinav Shrivastava. Learning to predict visual attributes in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13018–13028, 2021.

Vipin Pillai, Soroush Abbasi Koohpayegani, Ashley Ouligian, Dennis Fong, and Hamed Pirsiavash. Consistent explanations by contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10213–10222, 2022.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144, 2016.

Laura Rieger, Chandan Singh, William Murdoch, and Bin Yu. Interpretations are useful: penalizing explanations to align neural networks with prior knowledge. In *International conference on machine learning*, pages 8116–8126. PMLR, 2020.

Ribana Roscher, Bastian Bohn, Marco F Duarte, and Jochen Garcke. Explainable machine learning for scientific insights and discoveries. *Ieee Access*, 8:42200–42216, 2020.

Andrew Slavin Ross, Michael C Hughes, and Finale Doshi-Velez. Right for the right reasons: Training differentiable models by constraining their explanations. *arXiv preprint arXiv:1703.03717*, 2017.

Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv 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.

Ramprasaath R Selvaraju, Karan Desai, Justin Johnson, and Nikhil Naik. Casting your model: Learning to localize improves self-supervised representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11058–11067, 2021.

Oriane Siméoni, Gilles Puy, Huy V Vo, Simon Roburin, Spyros Gidaris, Andrei Bursuc, Patrick Pérez, Renaud Marlet, and Jean Ponce. Localizing objects with self-supervised transformers and no labels. *arXiv preprint arXiv:2109.14279*, 2021.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

Krishna Kumar Singh, Dhruv Mahajan, Kristen Grauman, Yong Jae Lee, Matt Feiszli, and Deepti Ghadiyaram. Don't judge an object by its context: Learning to overcome contextual bias. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11070–11078, 2020.

Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin Riedmiller. Striving for simplicity: The all convolutional net. *arXiv preprint arXiv:1412.6806*, 2014.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *International conference on machine learning*, pages 3319–3328. PMLR, 2017.

Yude Wang, Jie Zhang, Meina Kan, Shiguang Shan, and Xilin Chen. Self-supervised equivariant attention mechanism for weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12275–12284, 2020.

Matthew D Zeiler and Rob 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.

Zhaohui Zheng, Ping Wang, Wei Liu, Jinze Li, Rongguang Ye, and Dongwei Ren. Distance-iou loss: Faster and better learning for bounding box regression. In *The AAAI Conference on Artificial Intelligence (AAAI)*, 2020.

Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2921–2929, 2016.