

# Focus and Bias: Will It Blend?

Anna ARIAS-DUART <sup>a,1</sup>, Ferran PARÉS <sup>a</sup> and Víctor GIMÉNEZ-ÁBALOS <sup>a</sup>  
Dario GARCIA-GASULLA <sup>a</sup>

<sup>a</sup> *Barcelona Supercomputing Center (BSC)*

## Abstract.

One direct application of explainable AI feature attribution methods is to be used for detecting unwanted biases. To do so, domain experts typically have to review explained inputs, checking for the presence of unwanted biases learnt by the model. However, the huge amount of samples the domain experts must review makes this task more challenging as the size of the dataset grows. In an ideal case, domain experts should be provided only with a small number of selected samples containing potential biases. The recently published Focus score seems a promising tool for the selection of samples containing potential unwanted biases. In this work, we conduct a first study in this direction, analyzing the behavior of the Focus score when applied to a biased model. First, we verified that Focus is indeed sensitive to an induced bias. This is assessed by forcing a spurious correlation, training a model using only cats-indoor and dogs-outdoor. We empirically prove that the model learnt to distinguish the contexts (outdoor vs indoor) instead of cat vs dog classes, so ensuring that the model learnt an unwanted bias. Afterwards, we apply the Focus on this biased model showing how the Focus score decreases when the input contains the aforementioned bias. This analysis sheds light on the Focus behavior when applied to a biased model, highlighting its strengths for its use for bias detection.

**Keywords.** Focus, explainability, feature attribution methods, evaluation metrics, bias, mosaics

## 1. Introduction

Neural networks have proven to be effective tools for image classification tasks [1]. However, their interpretation is still obscure. The most popular methods used to overcome this problem are post-hoc attribution methods, such as SmoothGrad [2], GradCAM [3], Layer-Wise Relevance Propagation (LRP) [4] or LIME [5]. These methods provide an attribution map representing the contribution of pixels towards the final prediction. In order to assess the reliability of these techniques, different approaches have been proposed in the literature. Evaluation metrics such as the Pointing Game [6] or the Region Perturbation [7]. Recent works exploit the use of grids to carry out these evaluations [8,9].

In this work, we analyze in depth the behavior of the Focus [8] score, particularly when applied to a biased model. To do so, we trained a model on biased data, enabling it to learn a spurious correlation we can quantify and control. Then we analyse the Focus behaviour when applied to this biased model. Notice that these spurious correlations are difficult to detect and validate using classic performance metrics (e.g., loss, accuracy), since these unwanted biases helps the model to learn and obtain correct predictions.

---

<sup>1</sup>Corresponding Author: Anna Arias-Duart; E-mail: anna.ariasduart@bsc.es.

## 2. Related Work

One of the powerful uses of the feature attribution methods is the detection of unwanted biases in datasets and models. However, deciding whether these spurious correlations are desirable or undesirable is for domain experts to say, following ethical and practical considerations. To enable experts to do this job, exploiting these attribution maps, we need to provide them with useful and reliable information.

Some work has been done in this direction, Lapuschkin *et al.* [10] propose to reduce the explanations space provided to the domain experts through spectral clustering, so producing a reduce set of clusters instead of thousands of explanations. Where these clusters aim to represent different classification strategies and then, these strategies are shown to domain experts for the final unwanted bias detection stage. Similarly, this work [8] proposes a methodology where Focus is used to reduce the number of explanations to a subset that highlights potential unwanted biases, hence considerably reducing the search space to be assessed by domain experts for finding unwanted biases. However, the authors in [8] use models in which they have no control over the actual bias, limiting the reliability of the results. In our work, we go beyond, applying the Focus on a model to which we induce an unwanted bias.

The Focus metric involves three elements: a feature attribution method, a trained classification model and a set of mosaic samples. Each mosaic is composed of images of different classes, some of them from a *target class*, the specific class the explainability method is expected to explain. For example, if the *target class* of a given mosaic is the *dog* class, at least one of the images within the grid of the mosaic must belong to the *dog* class (see Figure 3 for examples of  $2 \times 1$  mosaics). Intuitively, if we ask a feature attribution method for the explainability of the *target class* on a mosaic, the Focus metric will measure the proportion of attribution lying on the *target class* squares, with respect to the total mosaic attribution. A random (uniform) attribution obtains a Focus equal to the proportion of squares.

In this context, we analyze the Focus behavior when applied to a model where the spurious correlation learnt by the model is known, verifying its potential use as a tool for bias detection.





## 3. Building a biased model

To apply the Focus on a biased model, we first need a biased dataset to learn from. In this section, we first explain how we created the biased dataset §3.1. Then, we introduce the training configurations §3.2. And last but not least, we perform some sanity checks to confirm that indeed we managed to introduce a spurious correlation into the model §3.3.



**Figure 1.** Examples of indoor/outdoor images: (a) cat-indoor (b) cat-outdoor (c) dog-indoor (d) dog-outdoor.

**Table 1.** Contexts included in each category for the Visual Genome dataset. The first and second column corresponds to the cat-outdoor and cat-indoor category. And the third and fourth column to the dog-outdoor and dog-indoor category respectively.



			
car, fence, grass, roof, bench, bird, house	speaker, computer, screen, laptop, computer mouse, keyboard, monitor, desk, sheet, bed, blanket, remote control, comforter, pillow, couch, books, book, television, bookshelf, blinds, sink, bottle, faucet, towel, counter, curtain, toilet, pot, carpet, toy, floor, plate, rug, food, table, box, paper, suitcase, bag, container, vase, shelf, bowl, picture, papers, lamp, cup, sofa	house, grass, horse, fence, cow, sheep, dirt, car, motorcycle, truck, helmet, snow, flag, boat, rope, trees, frisbee, bike, bicycle, sand, surfboard, water, fire hydrant, pole, skateboard, bench, trash can	screen, shelf, desk, picture, laptop, remote control, blanket, bed, sheet, lamp, books, pillow, curtain, container, table, cup, plate, food, box, rug, floor, cabinet, towel, bowl, television, carpet, sofa

### 3.1. Dataset creation

The creation of this dataset is motivated by the need to have control over some of the dataset biases. To do so, we use the MetaShift [11] to induce a correlation that we can quantify and control. This work clusters the images according to metadata. An annotated graph is created where each node represents a class in a specific context, for example *dog frisbee*. The distance between nodes represents the similarity between those contexts: *dog frisbee* will be closer to *dog grass* than *dog books*. The more contexts are shared within a class, the closer the nodes will be. Using the construction proposed by [11] we create a dataset composed of two classes (cat and dog) with two subclasses (indoor and outdoor), see Figure 1 for details.

We built the dataset with images from two well-known datasets, both providing contextual information: the Common Objects in Context (COCO) dataset [12] and the Visual Genome dataset [13]. Tables 1 and 2 show the exact contexts used for the construction of the indoor and outdoor subclasses for both datasets, the Visual Genome dataset and the COCO dataset respectively.

**Table 2.** Contexts included in each category for the COCO dataset. The first column corresponds to the outdoor contexts and the second to the indoor ones.

	
bicycle, car, motorcycle, airplane, bus, train, truck, boat, traffic light, fire hydrant, stop sign, parking meter, bench, frisbee, skis, snowboard, sports ball, kite, baseball bat, baseball glove, skateboard, surfboard, tennis racket	bottle, wine glass, cup, fork, knife, spoon, bowl, chair, couch, potted plant, bed, dining table, toilet, tv, laptop, mouse, remote, keyboard, cell phone, microwave, oven, toaster, sink, refrigerator, book, clock, vase, scissors, teddy bear, hair drier, toothbrush

### 3.2. Model

Next, we train a model using only samples from cats-indoor and dogs-outdoor. In this way, we introduce a spurious correlation, which could in fact occur in a real scenario: dog-outdoor images are more likely than cat-outdoor images.

For training the model, we use a total of 1,060 images per class (cats-indoor vs dogs-outdoor). Where 960 images per class were used for training and 100 for validation. We use the ResNet-18 [14] architecture, the AMSGrad [15] to optimize weights and we perform data augmentation during training: random rotation ( $[-30, 30]$  degrees), random crop and random horizontal flip with a chance of 50%. We reach a mean per class accuracy on the validation set of 87%, corresponding to the model with the minimum validation loss. From here we will call this model: the *biased model*.

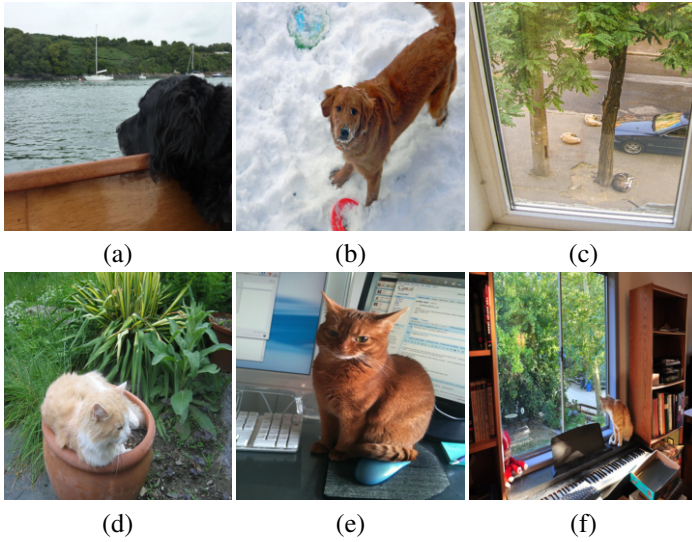
For comparison purposes, we also train a model avoiding the context correlation. We use the same training size (1,060 images per class) but in this case both cats and dogs will be equally present in both contexts 50% outdoors and 50% indoors. We reach a mean per class accuracy on the validation set of 60.5%, using the model with the minimum validation loss. Notice that the performance obtained is much lower, indicating that the induced context served as a successful shortcut to the model. Without this added bias, the high variability (different breeds) as well as the low quality (misabeled samples or partially occluded animals) of the dataset, limits the performance of the model which fails to learn to distinguish the two classes robustly. From now on we will refer to this second model as the *non-biased model*. Both trainings are performed in a single computing node of the CTE-Power9 cluster at the Barcelona Supercomputing Center, with the following characteristics:

- 2 × IBM Power9 8335-GTH @ 2.4GHz (20 cores and 4 threads/core).
- 4 × GPU NVIDIA V100 (Volta) with 16GB HBM2.

### 3.3. Sanity checks

To prove that the previous model, trained for the cats (indoor) and dogs (outdoor) classification task, is biased indeed (i.e., it has managed to learn the context instead of cat and dog characteristic patterns), we perform the following experiment. Starting from the hypothesis that images predicted with low probability, or that are predicted as the opposite class (in the case of a binary classification problem) are likely to be those that have patterns of the opposite class, we selected the three dog images with the lowest prediction and the three worst cat image predictions, see Figure 2.

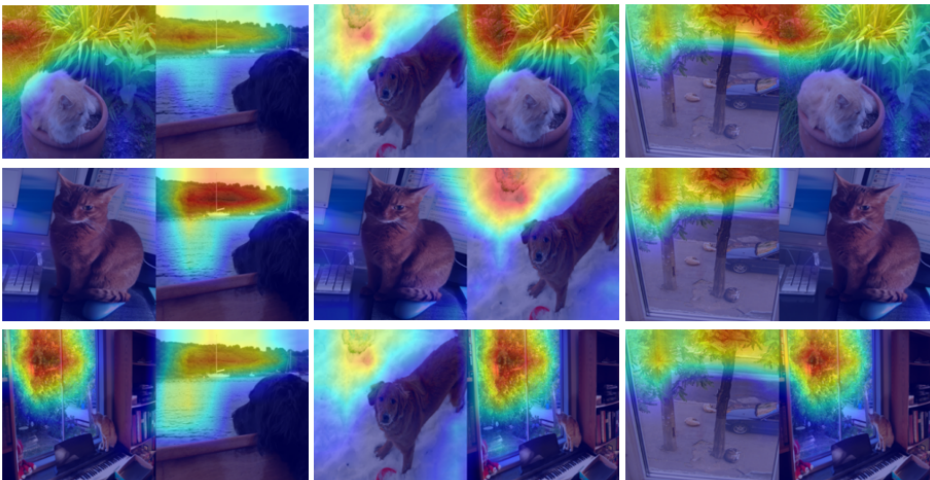
Before continuing with the hypothesis evaluation, it is worth mentioning how the samples predicted with least certainty significantly differ between cats and dogs. While for dogs, the lowest probability corresponds to 56.58% and the third lowest to 82.11% (both of which account for a correct classification in a binary problem), for cats these probabilities drop to 0.38% the lowest, and the third lowest to 47.29%. As shown in Figure 2 (and mentioned before) the worst predicted cat sample seems like a labeling mistake (labeled as cat indoor when it seems to be cat outdoor). We do not correct this mistake for the sake of methodological consistency. These results show a higher performance when classifying outdoor-dogs than indoor-cats, suggesting that the model has learnt to focus more on outdoor than indoor patterns. This may be due to the fact that outdoor patterns



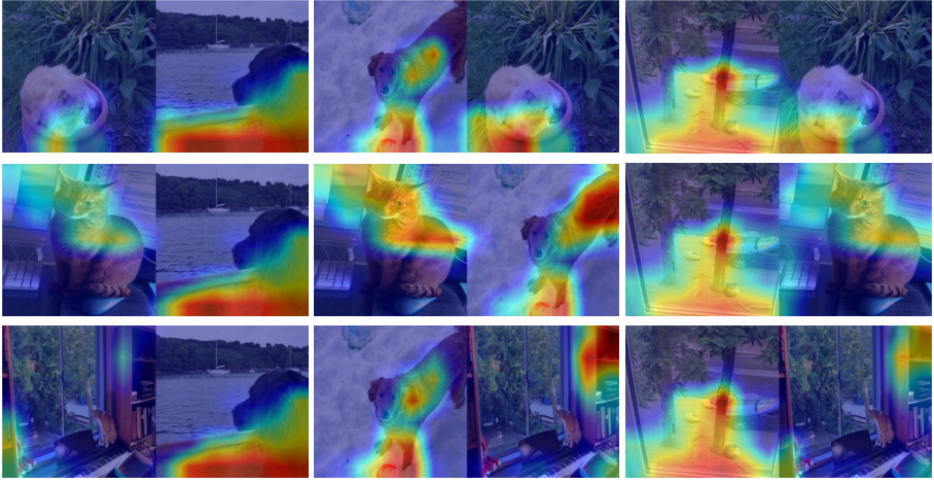
**Figure 2.** Examples of the worst predictions of the validation images set. Worst dog predictions: (a) dog: 0.5658, (b) dog: 0.7948 and (c) dog: 0.8211. Worst cat predictions: (d) cat: 0.0038, (e) cat: 0.4627 and (f) cat: 0.4729.

are less variant and more frequent, being a perfect visual pattern to discriminate between both classes.

Following the previous hypothesis: images predicted with a low probability are those that most likely contain a pattern of the opposite class. We build a set of  $2 \times 1$  mosaics by combining those pairs of images (cats vs dogs), in order to apply a feature attribution method on top of them. Notice that the use of feature attribution methods on top of the mosaics enhances the detection of shared biases (term introduced in [8]). The shared biases are characteristic patterns of one class that are present in another class.



**Figure 3.** Feature attribution maps obtained by GradCAM on the *bias model* (the dog being the *target class*).



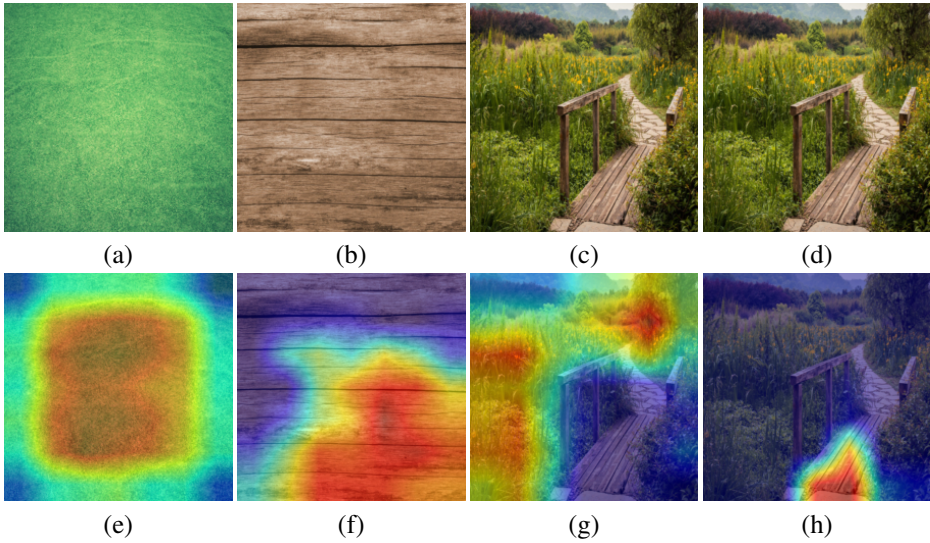
**Figure 4.** Feature attribution maps obtained by GradCAM on the *bias model* (the cat being the *target class*).

For this experiment we use the GradCAM attribution method. Results for both *target classes* are shown in Figure 3 and Figure 4. In all mosaics with the *target class* being the dog (see Figure 3), the GradCAM attribution focuses on areas where trees, leaves, or plants are present. Regardless of these patterns appearing in the cats squares or in the dog squares. Based on this, we hypothesize that the model has learnt to detect vegetation, instead of discriminating between cats and dogs. Indeed, it seems reasonable to think that most dogs in an outdoor context will be on meadows, fields or mountains (with a prominent presence of vegetation), while indoor cats will lack such pattern. This situation would have made it easier for the model to distinguish between dogs and cats, by only learning the green context instead of learning the characteristic patterns of these two mammals.

Similarly, the attribution in Figure 4, with the cat being the *target class*, falls on the wood or the brown areas (see for example first column of Figure 4). This pattern, although to a lesser extent than the vegetation, seems to be learnt by the model as a characteristic pattern of the cat class.

In order to corroborate that the model has learnt to identify vegetation as a characteristic pattern of the dog class and the wood as characteristic of the cat class, we perform another sanity check. We fed the model with the hand selected images shown in Figure 5, obtained from external sources. Image (a) is an image of only grass, which is predicted as a dog with a probability of 99.98%. On the contrary, Image (b) is a wood image which is predicted as a cat with a probability of 96.30%. In the case of Image (c), both patterns are present, although the green pattern is more prominent. This image is predicted as a dog with a probability of 99.44%. Notice how the attribution, being the *target class* the dog class (see Image (g)), falls on the green part around the path. However, when we ask for the attribution of the cat class (see Image (h)), the relevance focuses on the wooden bridge.

These results validate our hypothesis: vegetation is the main pattern learnt by the model as characteristic of the dog class and the wood pattern is learnt as characteristic of the cat class. At this point, we can confirm that the model is clearly skewed, it has learnt



**Figure 5.** First row: hand selected samples predicted by the *biased model* as (a) dog: 0.9998 (b) cat: 0.9630 (c) and (d) dog: 0.9944. Second row: feature attribution maps obtained by GradCAM for the images of the first row being the *target class*: (e) the dog class (f) the cat class (g) the dog class and (h) the cat class.

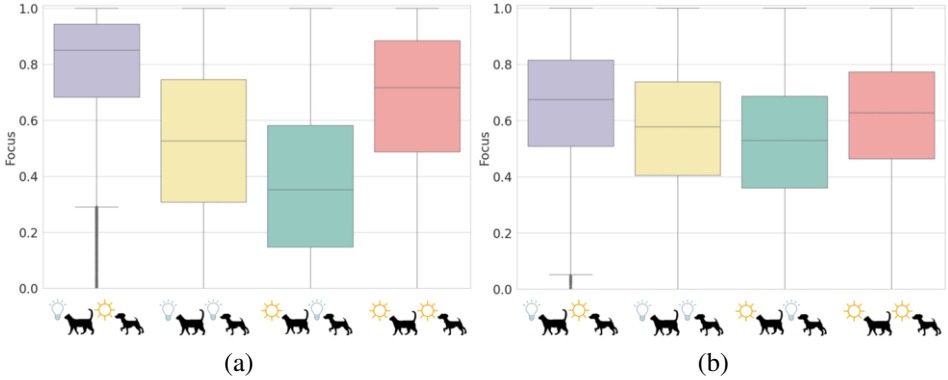
to differentiate the two classes mainly by context and not by animal, and furthermore, we are aware of the principal patterns enabling such distinction.

#### 4. Focus on a biased model

To evaluate the performance of the Focus score on a biased model we need three elements: a feature attribution method, a trained model and a set of mosaics. As an explainability method we use GradCAM, the one obtaining the best results for Focus [8]. As a trained classification model we use the ones introduced in §3.2 (the *biased model* and the *non-biased model*). Finally, for the mosaics, we build four sets of  $2 \times 1$  mosaics, following all possible combinations. Notice each set contains the same amount of mosaics (10,000):

1. **cat-indoor vs dog-outdoor:** Combines 100 cat-indoor images and 100 dog-outdoor images. Note that this set follows the same distribution used for training the *biased model*.
2. **cat-indoor vs dog-indoor:** Combines 100 cat-indoor images and 100 dog-indoor images.
3. **cat-outdoor vs dog-indoor:** Combines 100 cat-outdoor images and 100 dog-indoor images. Note that this set corresponds to a distribution complementary to the one used for training the *biased model*.
4. **cat-outdoor vs dog-outdoor:** Combines 100 cat-outdoor images and 100 dog-outdoor images.

Note that none of these sets corresponds to the distribution used for training the *non-biased model* in which samples of all sets are used (cats and dogs equally sampled from indoor and outdoor contexts). At this point we can now compute the Focus obtained by



**Figure 6.** Each box plot shows the Focus distribution for a different validation set (evaluating 10,000 mosaics per set). The purple box plots correspond to the cat-indoor and dog-outdoor set (Set 1). The yellow box plots correspond to the cat-indoor and dog-indoor set (Set 2). The green box plots to the cat-outdoor and dog-indoor set (Set 3). And the red box plots to the cat-outdoor and dog-outdoor set (Set 4). (a) Focus distributions obtained by GradCAM on the *biased model* (b) Focus distributions obtained by GradCAM on the *non-biased model*.

each of the two models on each of the four mosaic sets. The resulting Focus distributions (including the 10,000 samples per set) are shown in Figure 6.

In the experiments with the *biased model*, the highest Focus is expected to be obtained with the Set 1, since the images within this set follow the same distribution in which the model has been trained. On the other hand, the Focus obtained with the Set 3, should be the lowest, since the images correspond to the completely inverse distribution. In this case, the mean Focus is expected to be between 0 and 0.5 since the learnt biases may be found on the the non *target class* squares.

In the experiments with the *non-biased model*, we expect the Focus distributions to be similar to one another. The training distribution of this model avoid biases regarding indoor and outdoor, which should prevent the model from focusing on these properties. Thus, if the context is not a factor, the four sets become analogous.

As seen in Figure 6, results follow our hypothesis. The context (indoor/outdoor) plays a significant role in the *biased model*, and have a much weaker impact on the results of the *non-biased model*. For the *biased model*, a mean Focus greater than 0.8 is obtained when using the same context as in training (Set 1, see first box plot in Figure 6 (a)). However, when the complementary distribution is used, Set 3, the mean Focus falls below 0.4. As hypothesized before, this low Focus is most likely due to the model finding patterns in the image of the opposite *target class*. Finally, the two sets having at least one correct context (Set 2 and Set 4) obtain a mean Focus in between the two mentioned above (see the second and the fourth box plot in Figure 6 (a)).

We hypothesize that a significant amount of label noise is found (particularly in the cat outdoor class, incorrectly labeling indoor cat images as outdoor samples). This would explain the fact that outdoor cats and dogs (red box plot of Figure 6 (a)) obtains a higher Focus than indoor cats and dogs (yellow box plot of Figure 6 (a)) as well as why the inverse distributed set (green box plot of Figure 6 (a), mean Focus of 0.3532) is not the complementary of the equally distributed set (purple box plot of Figure 6 (a), mean Focus of 0.8507).

In contrast, the Focus distributions obtained with the *non-biased model* have a mean Focus close to each other. The mean Focus obtained with Set 1 is still the highest, as



shown in Figure 6 (b), and the mean Focus obtained with Set 3 is slightly the lowest. This is likely to be caused by label noise induced by the natural predominance of cats to be indoor, and of dogs to be outdoor.

## 5. Conclusions

In this paper we analyze the behavior of Focus when applied to a biased model. To do so, we train a model to classify cats and dogs, to which we induce a correlation: we only use cats-indoor and dogs-outdoor. In this way, we force the model to learn a bias, in this case the context. To verify that this model is indeed biased, we perform a set of sanity checks. For that we use an explainability method (GradCAM) on top of mosaics. The nature of mosaics allows us to easily identify the shared bias found within the model: the vegetation patterns were learnt by the model as characteristic of the dog class, while brown and wood patterns are learnt as characteristic of cat. We use this *biased model* to analyze the behavior of the Focus when applied to the biased setting. For baseline we use a *non-biased model*. To perform this experiment, we use 4 mosaic sets: cat-indoor vs dog-outdoor (Set 1), cat-indoor vs dog-indoor (Set 2), cat-outdoor vs dog-indoor (Set 3) and cat-outdoor vs dog-outdoor (Set 4). Our findings show how the presence of a shared bias is clearly reflected in the Focus distribution. The Focus decreases when the context learnt by the model is present in both classes within the mosaics. This shows the potential of the Focus, together with the mosaic structure, for the detection of unwanted biases in datasets and models.

## 6. Future Work

In this work, we empirically proved how the Focus is sensitive to the presence of bias in the model. However, it remains as future work to design a methodology to construct a reduced set of mosaics (selecting a pair of images) that highlight potential biases in the model. The key idea would be to provide to the domain experts the smallest number of mosaics containing as many model biases as possible. Domain experts would use such set of mosaics for the subsequent inspection, detection and classification between desirable and undesirable biases.

Our current method to select mosaics consist in selecting images with lowest prediction of its corresponding class (see §3.3) with the idea of containing characteristic patterns of the other class. However, in a non-binary problem, an image with low prediction score for a class may contain patterns from any of the other resting classes. This casuistic multiplies the number of mosaics that will be provided to the domain expert, thus increasing the complexity of their task.

An interesting case would be a mosaic with high accuracy and low Focus. On one side, the high accuracy would mean that the mosaic contains strong evidences of the *target class* while, on the other side, the low Focus score would mean that such evidences are shared between mosaic images, belonging and non-belonging to the *target class*. This remains as future work.

## Acknowledgements

This work is supported by the European Union – H2020 Program under the “INFRAIA-01-2018-2019 – Integrating Activities for Advanced Communities”, Grant Agreement n.871042, “SoBigData++: European Integrated Infrastructure for Social Mining and Big Data Analytics” and by the Departament de Recerca i Universitats of the Generalitat de Catalunya under the Industrial Doctorate Grant DI 2018-100.

## References

- [1] Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*. 2012;25.
- [2] Smilkov D, Thorat N, Kim B, Viégas F, Wattenberg M. Smoothgrad: removing noise by adding noise. *arXiv preprint arXiv:170603825*. 2017.
- [3] Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-cam: Visual explanations from deep networks via gradient-based localization. In: *Proceedings of the IEEE international conference on computer vision*; 2017. p. 618-26.
- [4] Bach S, Binder A, Montavon G, Klauschen F, Müller KR, Samek W. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLoS one*. 2015;10(7).
- [5] Ribeiro MT, Singh S, Guestrin C. ” 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*; 2016. p. 1135-44.
- [6] Zhang J, Bargal SA, Lin Z, Brandt J, Shen X, Sclaroff S. Top-down neural attention by excitation backprop. *International Journal of Computer Vision*. 2018;126(10):1084-102.
- [7] Samek W, Binder A, Montavon G, Lapuschkin S, Müller KR. Evaluating the visualization of what a deep neural network has learned. *IEEE transactions on neural networks and learning systems*. 2016;28(11):2660-73.
- [8] Arias-Duart A, Parés F, Garcia-Gasulla D, Gimenez-Abalos V. Focus! Rating XAI Methods and Finding Biases. *arXiv*; 2021. Available from: <https://arxiv.org/abs/2109.15035>.
- [9] Rao S, Böhle M, Schiele B. Towards Better Understanding Attribution Methods. In: *35th IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE; 2022. .
- [10] Lapuschkin S, Wäldchen S, Binder A, Montavon G, Samek W, Müller KR. Unmasking Clever Hans predictors and assessing what machines really learn. *Nature communications*. 2019;10(1):1-8.
- [11] Liang W, Zou J. MetaShift: A Dataset of Datasets for Evaluating Contextual Distribution Shifts and Training Conflicts. In: *International Conference on Learning Representations*; 2021. .
- [12] Lin TY, Maire M, Belongie S, Hays J, Perona P, Ramanan D, et al. Microsoft coco: Common objects in context. In: *European conference on computer vision*. Springer; 2014. p. 740-55.
- [13] Krishna R, Zhu Y, Groth O, Johnson J, Hata K, Kravitz J, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*. 2017;123(1):32-73.
- [14] He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*; 2016. p. 770-8.
- [15] Reddi SJ, Kale S, Kumar S. On the convergence of adam and beyond. *arXiv preprint arXiv:190409237*. 2019.