Focus! Rating XAI Methods and Finding Biases

Anna Arias Duart*†, Férran Parés*, Dario Garcia-Gasulla*
*Barcelona Supercomputing Center, Barcelona, Spain
†Universitat Politècnica de Catalunya, Barcelona, Spain
E-mail: {anna.ariasduart, ferran.pares, dario.garcia}@bsc.es

Keywords—Explainable AI, Bias detection, Image classification

I. Extended Abstract

Explainability has become a major topic of research in Artificial Intelligence (AI), aimed at increasing trust in models such as Deep Learning (DL) networks. However, trustworthy models cannot be achieved with explainable AI (XAI) methods unless the XAI methods themselves can be trusted.

To evaluate XAI methods one may assess interpretability, a *qualitative* measure of how understandable an explanation is to humans [1]. While this is important to guarantee the proper interaction between humans and the model, interpretability generally involves end-users in the process [2], inducing strong biases. In fact, a qualitative evaluation alone cannot guarantee coherency to reality (*i.e.,* model behavior), as false explanations can be more interpretable than accurate ones. To enable trust on XAI methods, we also need *quantitative* and objective evaluation metrics, which validate the relation between the explanations produced by the XAI method and the behavior of the trained model under assessment.

In this work we propose a novel evaluation score for *feature attribution* methods, described in §I-A. Our input alteration approach induces in-distribution noise into samples, that is, alterations on the input which correspond to visual patterns found within the original data distribution. To do so we modify the context of the sample instead of the content, leaving the original pixels values untouched. In practice, we create a new sample, composed of samples of different classes, which we call a *mosaic image* (see examples in Figure 2). Using *mosaics* as input has a major benefit: each input quadrant is an image from the original distribution, producing blobs of activations in each quadrant which are consequently coherent. Only the pixels forming the borders between images, and the few corresponding activations, may be considered out of distribution.

By inducing in-distribution noise, *mosaic images* introduce a problem in which XAI methods may objectively err (focus on something it should not be focusing on). On those composed mosaics we ask a XAI method to provide explanation for just one of the contained classes, and follow its response. Then, we measure how much of the explanation generated by the XAI is located on the areas corresponding to the target class, quantifying it through the *Focus* score. This score allows us to compare methods in terms of explanation precision, evaluating the capability of XAI methods to provide explanations related to the requested class. Using *mosaics* has another benefit. Since the noise introduced is in-distribution, the explanation errors identify and exemplify biases of the model. This facilitates the elimination of biases in models and datasets, potentially resulting in more reliable solutions. We illustrate how to do so in §I-C.

A. The Focus metric

When a *feature attribution* method is applied to an image to explain the model's prediction regarding a chosen class, it typically produces a map from pixels to real values, referred to as relevance. To formalize mosaics, and later *Focus*, let us define a dataset D composed by a set of images $I = \{img_1, img_2, ..., img_N\}$ and a set of classes $C = \{c_1, c_2, ..., c_K\}$, where N is the number of total images and K is the number of total classes. Every image in I has assigned a unique class from C: $c(img)$. From here we build a set of mosaics $M = \{m_1, m_2, ..., m_J\}$ where J is the total number of mosaics in M. A mosaic m is composed by four images $m = \{img_1, img_2, img_3, img_4\}$ and characterized by a target class $tc = c(m)$, the specific class the XAI method is expected to explain. While two images of the mosaic belong to the target class $c(img_1) = c(img_2) = c(m)$, the other two are randomly chosen among the rest of classes $c(img_3) \neq c(m); c(img_4) \neq c(m)$. Mosaics are implemented as two by two, non-overlapping grid, with the position of each image being random.

The *Focus* metric estimates the reliability of XAI method's output as the probability of the sampled pixels lying on an image of the target class of the mosaic $c(m)$. This is equivalent to the proportion of positive relevance lying on those images:

$$F_{\mathcal{A},\theta}(m) = \frac{R_{c(m)}(img_1) + R_{c(m)}(img_2)}{R_{c(m)}(m)} \quad (1)$$

where $R_c(r)$ is the sum of positive relevance toward class c on the region of the mosaic r. This probability can be interpreted as a precision of the relevance. In an sort of eye-tracking game, the *Focus* metric asks to the XAI method "*Why does mosaic m belong to class $c(m)$?*" on a mosaic m which contains both samples belonging and not belonging to the target class $c(m)$. Given the previous question and a good underlying model, a reliable *feature attribution* method should be able to concentrate most of its explanation relevance on the two appropriated images of the mosaic (img_1 and img_2).

B. Evaluation of XAI methods

We evaluate GradCAM, LRP, SmoothGrad, LIME, GradCAM++ and IG, using three architectures (AlexNet, VGG16 and ResNet-18) and four target datasets (Dogs vs. Cats, MAMe, MIT67 and ImageNet). Figure 1 shows an example of the *Focus* distribution obtained for the MAMe dataset experiments.

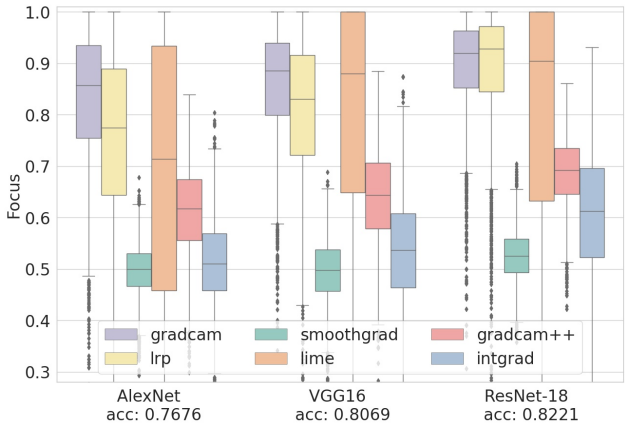

Fig. 1. *Focus* distribution boxplot for different XAI methods applied to models trained for the MAMe dataset.

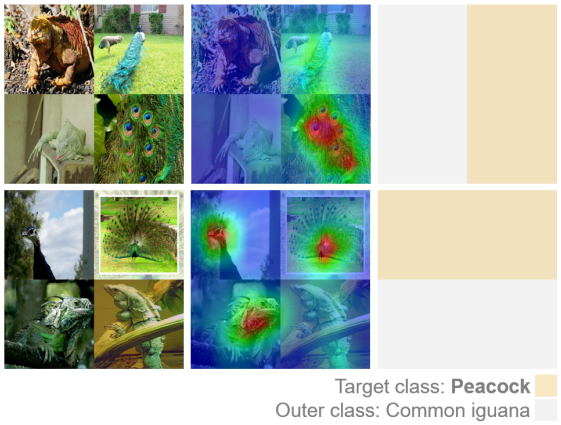

Fig. 2. GradCAM explanations obtained on the ResNet-18 trained with ImageNet. Two examples of mosaics are shown in the first column. The second column shows the corresponding GradCAM explanations for the target class. The third column specifies the positions of the classes within the mosaic. The target class is the *Peacock* class and the outer class is the *Common iguana* class. The example above obtains a high *Focus* score (0.8176) and the one below a lower one (0.4940).

When applied to SmoothGrad or IG, *Focus* finds these methodologies as quasi-random in its explanations with respect to the model. On the contrary, LRP and GradCAM are both found to be consistently reliable methods. GradCAM performs well on all experiments conducted, even when the underlying model is not particularly well fit to the task. LRP performs very well for high performing models, but it becomes more unreliable on less accurate models. This also seems to be the case of LIME, which suffers from an even larger variance. GradCAM++ performs better than random, but not as well as GradCAM and LRP.

C. Bias detection

We can also use the *Focus* to automate the bias identification in models and datasets. This is possible because mosaics induce in-distribution noise, where *Focus* errors directly correspond to visual biases of the model. The proposed procedure is as follows. First, for a better detection of biases between pairs of classes, we use mosaics with two classes. Therefore, in the mosaics used for this section, samples different from the target class actually belong to the same class: $c(img_3) = c(img_4) \neq c(m)$. We concentrate on the most relevant biases by finding the pairs of classes obtaining the lowest mean *Focus* in their joint mosaics. For each of these pairs we extract the mosaics with highest and lowest *Focus*, and present them to a human evaluator who must review the explanations produced. The role of the evaluator is to interpret the rationale behind the explanations (both correct and incorrect) and its degree of generalization for the task. Based on that assessment, corrective measures can be implemented.

To conduct this experiment we use the GradCAM method and the ResNet-18 architecture, a configuration which obtains a particularly robust *Focus*. An example is shown in Figure 2, divided in two rows: the top one corresponds to a high *Focus* and the bottom one to a low *Focus*. In this example, the model is able to correctly attribute relevance to the *Peacock* images on the upper mosaic, while, for the bottom mosaic, some of the relevance incorrectly fall on the head of the *Common iguana*. The fact that most of the incorrect relevance in the *Common iguana* falls in the subtympanic shield (the characteristic circle in its jowl) seems to be related with its visual similarity with the ocellus of the *Peacock* (the circular spot in the feathers). Notice the iguana's subtympanic shield is hardly visible in the top mosaic. After the identification of these biases, and an assessment of their impact, one could try to mitigate their relevance for the model. For example more images of the target class without the characteristic pattern found in the outer class could be added to the training set (*Peacocks* images where the ocellus is not visible).

D. Conclusion

With the aim of evaluating XAI methods in a quantitative manner, we introduce a novel metric—the *Focus*—to assess the faithfulness of a XAI method to the underlying model. We show the methodology to be consistent across tasks and architectures, providing strong empirical evidence of their performance. We introduce another application of *Focus*, using it for the identification and characterization of biases found in models. This empowers bias-management tools, in another small step towards trustworthy AI.

References

[1] L. H. Gilpin *et al.*, "Explaining explanations: An overview of interpretability of machine learning," in *2018 IEEE 5th International Conference on data science and advanced analytics (DSAA)*. IEEE, 2018, pp. 80–89.

[2] S. Mohseni *et al.*, "A multidisciplinary survey and framework for design and evaluation of explainable ai systems," *arXiv preprint arXiv:1811.11839*, 2018.

Anna Arias Duart obtained a Bachelor's degree in Telecommunications Technology and Services Engineering in 2015 from the Universitat Politècnica de València (UPV). In 2018 she completed the Double Diploma awarded by the UPV and Télécom ParisTech (Paris). She is currently a student of the Doctoral Program in Artificial Intelligence since 2019 at the Universitat Politècnica de Catalunya within the Industrial Doctorate Program in collaboration with SEAT, S.A.