ISSN 1870-4069

A Novel Data Augmentation Method based on XAI

Tonantzin Marcayda Guerrero Velázquez, Juan Humberto Sossa Azuela

Instituto Politécnico Nacional,
Centro de Investigación en Computación,
Laboratorio de Robótica y Mecatrónica,
México

{tmguerrerov, humbertosossa}@gmail.com

Abstract. Machine learning systems allow solving a wide variety of problems present both in industry and in everyday life. The increasing available information has made possible the training of these models and reaches a precision that surpasses the performance of a human being when carrying out the same task. However, the large amount of information necessary to carry out the training task of these models is not always available or difficult to obtain. That is why several methods are used to increase that information from existing training data; these methods are called data augmentation methods. The use of data augmentation techniques allows increment artificially the numbers of samples of a small training dataset, improving the performance of a machine learning model in tasks for which the information available is limited. This paper presents a novel data augmentation technique based on explainable artificial intelligence, where new training samples are created from the explanations generated using the explainability method described in [1]. This method generates explanations based on the classification of useful regions found in an image. Using this method, we can improve the model performance and its accuracy on the test dataset.

Keywords: XAI, CNN, data augmentation,explainability, machine learning.

1 Introduction

Data augmentation is a widely used technique to artificially create new samples information from a small training dataset. This data augmentation is done to obtain a greater range of features and information to feed the models we will use, for example, a CNN (Convolutional Neural Network) that is used for image classification or object recognition tasks. In order to obtain this new augmented dataset, talking about images, several techniques could be applied; some of the most common is based on changing the positioning or the color of the image.

Some of these techniques are: horizontal and vertical shift augmentation (HVA), horizontal and vertical flip augmentation (HVF), random rotation augmentation (RRA), random brightness augmentation (RBA), and random zoom augmentation (RZA). Machine learning models provide solutions to many problems with high precision, but this precision is could due to the quality of information available.

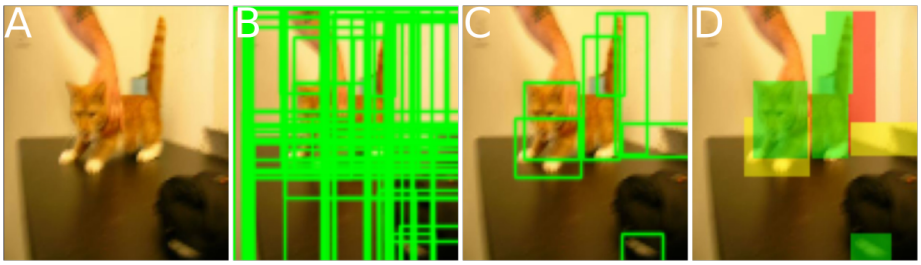

Fig. 1. Generation of the visual explanation of the prediction of an image classifier model. (A) Original image, (B) Set of Candidate regions, (C) Set of useful regions (D) Visual explanation.

However, some cases do not have much information, and the available data is limited, or it is not so easy to obtain; such is the case of the analysis of medical images [2]. Then, it is beneficial for these cases to apply data augmentation techniques. New data augmentation techniques have emerged to solve the problem of having small training dataset. Such is the case of the work described in [3] where they present a new technique based on randomly cutting out regions of four different images and putting them together to form a new image.

This new image will be part of the training dataset, and its label will be proportionally composed of the classes to which the cropped images belong. In [4] a technique called Smart Augmentation is described, which is based on creating a new neural network that learns how to generate new data during the training process of the base model. Another interesting technique is the one described in [5], which uses a set of functions to combine different input images to create a new image. This new image has an unrealistic appearance for a human being; however, according to the results presented by the authors, this technique generates good results for increasing data.

More recent work is that described in [6], where a technique is proposed that is based on randomly selecting a rectangular region of the image and erasing its pixels with random values during the training process. The author reports that with his methodology a reasonable improvement was obtained in the detection of objects as well as in the identification of people. In [7]a data augmentation method is proposed based on producing new information using the explanations generated by applying various existing explainability methods. Its objective is to try to align the predictions obtained from a model with the explanations resulting from the applied explanatory method.

2 Data Augmentation

Data Augmentation methods seek to artificially increase the original training dataset by applying different transformations to this data. In [8] two categories are described that encompass this type of transformation: geometric and photometric. Geometric transformations refer to those that alter the geometry of the image, moving each of its pixels in a new direction. On the other hand, the photometric or color transformations refer to those that alter the value of the RGB channels by changing each pixel value (r, g, b) to a new value (r', g', b') according to the predefined heuristics [8].

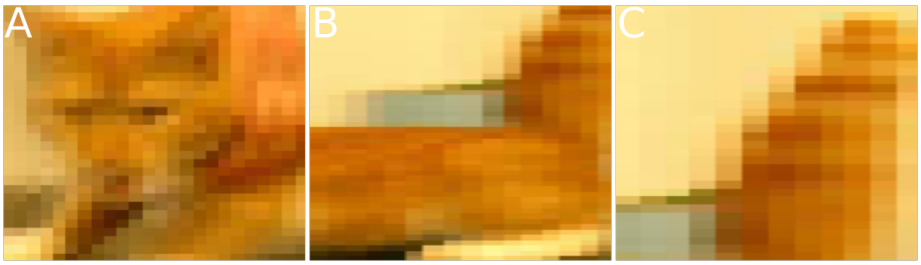

Fig. 2. Example of augmented data generated by applying the explainability method for Figure 1(A).

The data augmentation technique seeks to improve the performance of an image classifier where the amount of information in the training dataset is small. One of the problems with a small dataset is that the model does not generalize adequately for data from the test and validation set, and it is easy that presents a model overfitting problem. With the use of data augmentation, it is possible to reduce this overfitting [9].

Data augmentation is not the only one used to reduce the overfitting of a model. There are other techniques such as Dropout, Batch normalization, Transfer Learning, and Pretraining. However, unlike these techniques, data augmentation deals directly with the root of the problem, that is, the lack of data in the training dataset, as is show in [2].

3 Explainable Artificial Intelligence XAI

Today, machine learning models and deep learning models are widely used to solve complex problems and automate processes, both in industry and everyday life. These models reach surprising levels of precision and good performance, which have often exceeded the results obtained by a human being. The fact that they generate such good results can lead to some mistrust in them since, generally, these models are treated as a black box, of which only the inputs and outputs obtained when processing the information provided are known.

Because these models are used in critical areas such as medicine, medical care, autonomous vehicle management, credit approval, legal and justice issues, etc... , the need arose to create techniques that generate confidence in these models and the results they generate. These techniques are used in order to know the reason for their decisions and how they arrive at such precise results.

They are included in the so-called Explainable Artificial Intelligence (XAI) that arises from satisfying this need and seeks to generate methods that allow understanding the predictions of AI models as well as trying to justify their decisions. There are many explanation techniques that, when are applied to the models, allow us to understand the reason for their results. These methods can be used on different types of data, including images. The methods that provide a visual explanation of the prediction denote areas of the image regions that are of interest to the model when carrying out its prediction so that the explanation can be easily observed.

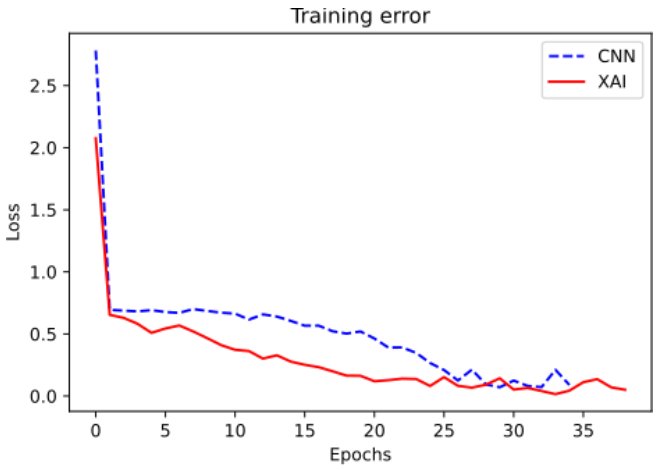

Fig. 3. Training error with and without data augmentation XAI.

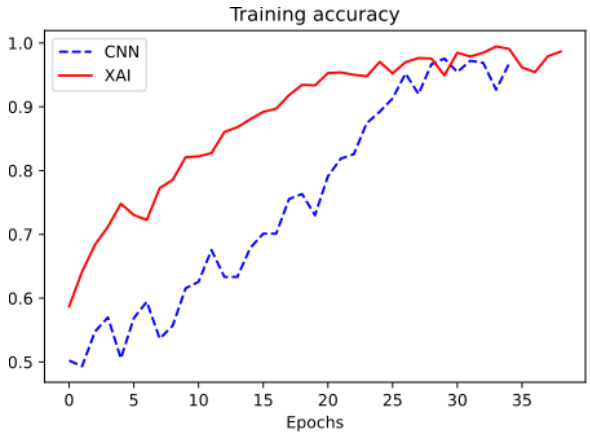

Fig. 4. Training accuracy with and without data augmentation XAI.

In the present work, a new Data Augmentation technique based on XAI is proposed. The augmented data is obtained from the results of the explanations obtained with a explainability method based on the classification of useful regions of an image. The implementation details of the proposed explainability technique are described in [1].

4 Proposed Method

The explainability method [1] in which this paper is based is used for image classifiers and obtains the visual explanation of the prediction of an image classifier model by identifying some regions of the input image (Figure 1(A)) that turn out to be most important for the prediction of the model, called useful regions.

Table 1. Results on the test dataset with and without data augmentation XAI.

Model	Training ACC	Validation ACC	Testing ACC	AUC
CNN	0.9278	0.7700	0.6000	0.786
XAI	0.9789	0.7800	0.6500	0.8371

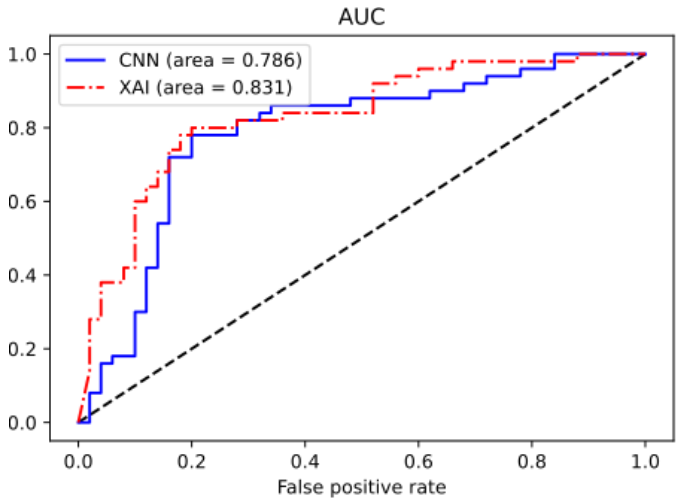

Fig. 5. Curves ROC comparison.

These regions are later statistically categorized, as significant, relevant, or futile and highlighted with green, yellow, and red colors, respectively, to get a visual explanation. The significant regions are marked in green, as can be seen in Figure 1(D). To obtain this explanation, the first thing that is done is a selective search for regions in the input image in order to obtain a set of candidate regions as show in Figure 1(B), so-called because it is not yet known whether they are important to the classifier.

For this reason, each of the regions of this set of candidate regions is evaluated with the classifier and subsequently subjected to statistical analysis to select the really important regions, which are called useful regions which are denoted in Figure 1(C). Finally, this set of useful regions is categorized and colored according to their importance in significant, relevant, or futile, as we can see in Figure 1(D). With this, the level of importance that each of these regions represents for the classifier model is denoted at the time of carrying out its prediction.

The details of this process are described in our work cited in [1]. Figure 1 shows the hole process to obtain a visual explanation for an image, using a Cat vs Dog classifier. The proposed Data Augmentation technique is based on this explainability method [1] and consists of adding to the original training dataset all those regions categorized as significant (black regions in Figure 1(D)) that result from the explanation obtained for each of the data in the training dataset that were correctly classified by the model.

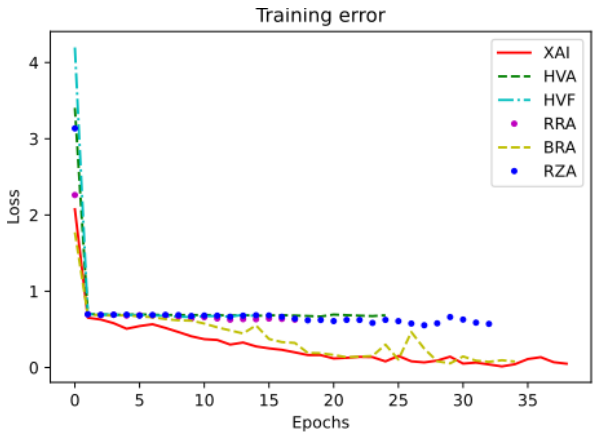

Fig. 6. Training error data augmentation methods comparison.

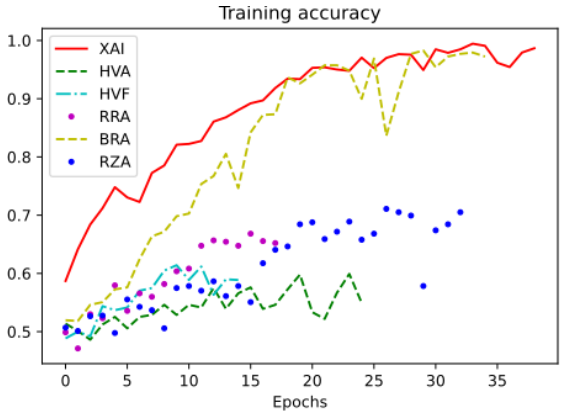

Fig. 7. Training Accuracy data augmentation methods comparison.

Figure 2 shows data that belongs to the augmented dataset generated by applying the explainability method to the original 100 x 100 pixels image. Then the created new samples are the regions identified with the explanation of the training dataset predictions and using the significant regions, it can be ensured that relevant information will be provided to the model, increasing the performance and accuracy of the model.

5 Results and Discussion

In the present work, we use the dataset Dogs vs. Cats that was taken from Kaggle [10], which contains a total of 25,000 images for the training dataset and 12,500 images for the test dataset. However, only a subset of 1000(100x100 pixels) images was taken from the total, in order to simulate a very small dataset.

Table 2. Comparison between the accuracies of the different methods.

Model	Training ACC	Validation ACC	Testing ACC	AUC
XAI	0.9789	0.7800	0.6500	0.831
HVA	0.6467	0.6900	0.6600	0.707
HVF	0.6289	0.6600	0.6200	0.685
RRA	0.6589	0.7000	0.6250	0.714
BRA	0.9644	0.8000	0.6450	0.846
RZA	0.7256	0.7800	0.6700	0.798

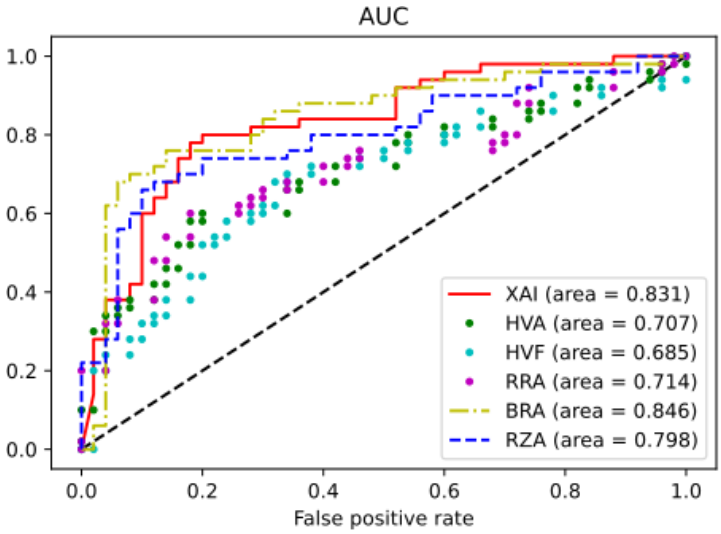

Fig. 8. AUC comparison.

Throughout this work, we use a CNN model with *Adam Optimization* and *categorical_crossentropy* as the loss function. From the 1000 images of the new small dataset, 90% are used for training and 10% for validation. Another different set of 200 images from the remaining 24000 from the original training dataset will be used for testing.

5.1 Data Augmentation XAI Based Method

After much effort, the best result we could get with our CNN model was accuracy of 0.9278 for training and 0.77 for validation, and using data augmentation with XAI as describe before we get an accuracy of 0.98 for training and 0.78 for validation. A comparison between the loss and accuracy evolution in training with and without data augmentation using XAI is shown in Figures 3 and 4. When we compare Figures 3 and 4, and according to the accuracy of the best model obtained, the data augmentation method using XAI proposed here slightly improves the model and its training.

Furthermore, if we use the AUC of the ROC curve as shown in Figure 5, the model trained used data augmentation is a better model, and it will be chosen in a real production environment where even 0.01 of improvement could make a difference. A better improvement is evident when we compare the results over the test dataset where the XAI method for data augmentation shows better accuracy as is shown in Table 1.

5.2 XAI Method Versus Known Methods

It is fair and necessary to compare the method against others commonly used. To do this, we use five different commonly used methods, horizontal and vertical shift augmentation HVA, horizontal and vertical flip augmentation HVF, random rotation augmentation RRA, random brightness augmentation RBA and random zoom augmentation RZA. In Figure 6, a comparison of the training error is shown where we can observe that the XAI method error is always lower than the others, and in Figure 7, we compare the training accuracy curve where the XAI method is also better. When we compare the AUC for the different methods, we found that the model created with the XAI method is one of the best methods, as shown in Figure 8. Table 2 shows a comparison between the accuracies obtained of the different methods of data augmentation on the different datasets.

6 Conclusions and Future Work

It was proven in this work that our proposed method for data augmentation based on XAI is efficient to obtain a better performance for the model than the one that is achieved without increasing data, this is without causing overfitting or reduction in the accuracy; it was also compared with other methods already known, demonstrating be one of the most outstanding. Our method can also be used in conjunction with any of the methods mentioned above to obtain better results, which will be left for future works.

Acknowledgment

Tonantzin Guerrero thanks CONACYT for the scholarship to undertake her doctoral studies.The authors thank the Instituto Politécnico Nacional for the economic support under projects SIP 20200630 and 20210788, and CONACYT under projects 65 (Fronteras de la Ciencia) and 6005 (FORDECYT-PRONACES).

References

1. Guerrero Velázquez, T. M., Sossa Azuela, J. H.: New explainability method based on the classification of useful regions in an image. Computación y Sistemas, vol. 25, no. 4 (2021) doi: 10.13053/cys-25-4-4049
2. Shorten, C., Khoshgoftaar, T. M.: A survey on image data augmentation for deeplearning. Journal of Big Data, vol.6, no. 1, pp. 1–48 (2019) doi: 10.1186/s40537-019-0197-0

3. Takahashi, R., Matsubara, T., Uehara, K.: Data augmentation using random image cropping and patching for deep CNNs. IEEE Transactions on Circuits and Systemsfor Video Technology, vol. 30, no. 9, pp. 2917–2931 (2019) doi: 10.48550/arXiv.1811.09030
4. Lemley, J., Bazrafkan, S., Corcoran, P.: Smart augmentation learning an optimal data augmentation strategy. IEEE Access, vol. 5, pp. 5858–5869 (2017) doi: 10.1109/ACCESS.2017.2696121
5. Summers, C., Dinneen, M. J.: Improved mixed-example data augmentation. In: IEEE Winter Conference on Applications of Computer Vision (WACV), pp. 1262–1270 (2019) doi: 10.48550/arXiv.1805.11272
6. Zhong, Z., Zheng, L., Kang, G., Li, S., Yang, Y.: Random erasing data augmentation. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, no. 7, pp. 13001–13008 (2020) doi:10.48550/arXiv.1708.04896
7. Li, R., Zhang, Z., Li, J., Sanner, S., Jang, J., Jeong, Y., Shim, D.: EDDA: Explanation-driven data augmentation to improve model and explanation alignment (2021) doi: 10.48550/ARXIV.2105.14162
8. Taylor, L., Nitschke, G.: Improving deep learning with generic data augmentation. In: IEEE Symposium Series on Computational Intelligence (SSCI), pp. 1542–1547 (2018) doi: 10.48550/ARXIV.1708.06020
9. Perez, L., Wang, J.: The effectiveness of data augmentation in image classification using deep learning (2017) doi: 10.48550/ARXIV.1712.04621
10. Microsoft. PetFinder.com: (Dogs vs.cats dataset)
11. Feng, S. Y., Gangal, V., Wei, J., Chandar, S., Vosoughi, S., Mitamura, T., Hovy, E.: A survey of data augmentation approaches for NLP (2021) doi: 10.48550/ARXIV.2105.03075
12. Montserrat, D. M., Lin, Q., Allebach, J., Delp, E. J.: Training object detection and recognition CNN models using data augmentation. Electronic Imaging, vol. 10, pp. 27–36 (2017) doi: 10.2352/ISSN.2470-1173.2017.10.IMAWM-163
13. Hernández-García, A., König, P.: Further advantages of data augmentation on convolutional neural networks. In: International Conference on Artificial Neural Networks. Springer, Cham, pp. 95–103 (2018) doi: 10.1007/978-3-030-01418-6_10
14. Kukačka, J., Golkov, V., Cremers, D.: Regularization for deep learning: A taxonomy (2017) doi: 10.48550/ARXIV.1710.10686
15. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.: Dropout: a simple way to prevent neural networks from overfitting. The journal of machine learning research, vol. 15, no. 56, pp. 1929–1958 (2014)
16. Ioffe, S., Szegedy, C.: Batch normalization: Accelerating deep network training by reducing internal covariate shift. In: International Conference on Machine Learning, pp. 448-456 (2015) doi: 10.48550/ARXIV.1502.03167
17. Molnar, C.: Interpretable Machine Learning. A guide for making black box models explainable (2020) https://christophm.github.io/interpretable-ml-book/.
18. Gunning, D.: Explainable Artificial Intelligence (XAI). DARPA (2017)