EXPLAINING DATA AUGMENTATION IN IMAGE CLASSIFICATION

*Shuai Wang, Nina Weng, Federico Emiliano Crespo Collazo**

Technical University of Denmark
Email: {s200108,s202997,s200161}@student.dtu.dk

ABSTRACT

Data augmentation has been heavily used as a mean of reducing overfitting and improve accuracy in Machine Learning problems. But the reasoning on why it does work has not been fully explored. Explainability, on the other hand, has become increasingly important in order to unravel black box algorithms. In this project, we will combine these two concepts to understand why Image Augmentation works. In this paper, we introduce a new method to quantitatively evaluate the similarity between explanation and ground truth (OV score) and prove that, at least for the performed experiments, using image augmentation not only accuracy but also the similarity between explanation and ground truth are improved.

***Index Terms*—** Explainability, data augmentation, image classification, LIME

1. INTRODUCTION

Data augmentation has been proved to be an efficient technique for improving performance in Machine Learning problems [1, 2]. As humans we have an intuition on why it could work, but could we explain how?

Explainability methods have become popular due to the need of explaining the black box algorithms to make them to be fully trustworthy [3]. This project will use explainability methods to further understand how image augmentation could help the learning process.

Previous researches rarely focus on using explainability to understand data augmentation. Chen et al. [4] proved that the explainability could be improved by data augmentation in the field of sentiment analysis. However, the mystery between data augmentation and prediction accuracy still remains unexplored.

In this article, firstly, in order to compare the explainability results, we introduce a method using the overlapping area of explanations and estimated ground truth to quantitatively evaluate how good the explanation result is. Based on that, we propose a framework for examining the results of both raw and augmented data, which includes the data augmentation process, model training, explainability analysis and the quantitative evaluation over the explanations. After this, two experiments about MNIST [5] and Dogs v.s. Cats classification [6] are carried out with the proposed framework. Finally, we analyse these expeirments trying to find a correlation in between data augmentation and explainability scores.

2. THEORETICAL BACKGROUND

Explainability, in this context, is highly related with black-box models, for which we are unsure about how everything works inside it. By exploring the explainability a human-like explanation on the descisions of the black box model is sought. In general, we are interested in explainability because it can help the developers to better improve the model. A typical instance here is classifying wolves and huskies[7], as almost all pictures of wolves have snow as background, the detector turn out to focus mainly on the snow rather than dog face.

This project uses LIME (Local Interpretable Model-agnostic Explanations)[7] which is an explanation technique that explains the prediction of any classifier by learning an interpretable model locally around the prediction.

An interpretable representation would vary with the type of data that we are working with. In our case, for images, it represents the presence/absence of superpixels (contiguous patch of similar pixels). Data augmentation [8] in data analysis are techniques used to increase the amount of data by adding slightly modified copies of already existing data or newly created synthetic data from existing data. It acts as a regularizer and helps reduce overfitting when training a machine learning model. It is closely related to oversampling in data analysis. Geometric transformations, flipping, color modification, cropping, noise injection and random erasing are used to augment image in deep learning.

3. METHODOLOGY

In this section, the general framework of the proposed method is introduced, followed by the constructed method (overlapping score) for quantitative evaluation over explanation re-

*Thanks to our supervisor Lars Kai Hansen for his guidance and great advice.

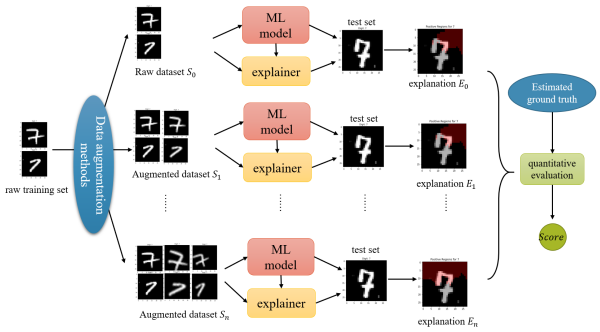


Fig. 1. Framework of the proposed method. Training sets are firstly extended with or without data argumentation, then the explanation of test sets are being quantitatively evaluated based on defined estimated ground truth.

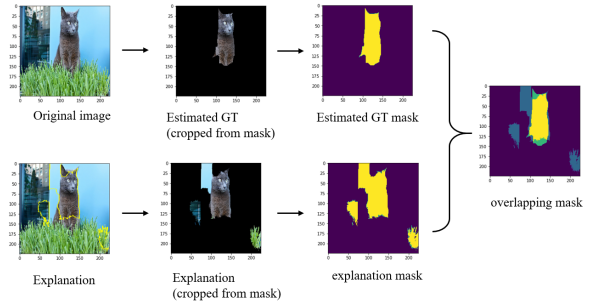


Fig. 2. The illustration of quantitative evaluation for Experiment 4.2.

sults. Two experiments about handwritten digits classification and pets classification are conducted in the next section to better illustrate how this framework works.

In order to explore how data augmentation improves the accuracy in the Machine Learning task, the framework which showed in Figure 1 is established. Firstly, after the data is split into the training set and the test set, one or multiple augmented datasets are built based on the original one. The training process is performed on all training sets followed by the explanability methods. After that, the given explanations are quantitatively evaluated based on the estimated ground truth(GT).

In the whole process, the two difficulties we are faced with are defining the estimated GT for explanation results, and evaluating the explanation quantitatively. The lack of GT for explanation is the bottleneck for evaluating the explanation results[9]. The principles of defining the estimated GT are: 1) the GT matches human's intuition and understanding; 2) the qualitative evaluation should be able to adjust according to the precision of estimated ground truth.

For evaluating the explanation results, the method proposed is inspired by the mask provided by explanation methods. In LIME, the explanations are showed as a mask, which indicates the area contributing the most to the prediction (the second row in Figure 2). Therefore, an intuitive thought would be the more overlapping area of estimated GT and the explanation mask have in common, the better the explanation is. To quantificationally evaluate prediction explanations, we propose the overlapping(OV) score defined in the following Equation 1 and 2, where M_{OV}, M_{Exp} and M_{GT} represents the 0-1 mask-like matrix. The score for overlapping area $score_{OV}$ contains two terms corresponding to the percentage of overlapping area in GT mask and explanation mask respectively. The p here represents the proportion for each term, which could be revised based on how precise the estimated GT is. $\circ$ represents element-wise multiplication in Eq. 1.

$$M_{OV} = M_{Exp} \circ M_{GT} \tag{1}$$

$$score_{OV} = p \times \frac{\sum M_{OV}}{\sum M_{Exp}} + (1-p) \times \frac{\sum M_{OV}}{\sum M_{GT}} \tag{2}$$

4. EXPERIMENT

4.1. Experiment 1: Handwritten Digits Classification

4.1.1. Dataset and Implementation details

MNIST[10] dataset is used in the first experiment, which contains 70,000 labeled images in total. With different kinds of data augmentation methods, there are all together 14 groups(runs) in this experiment which are shown in the y-axis of Figure 3(a). The first eight groups are the original sampled dataset and its augmented sets, and the next three groups consist of different number of repetitions of the sampled dataset, and the last three groups are differently sized subsets of the original dataset. The details of the 14 groups are shown in Appendix A.

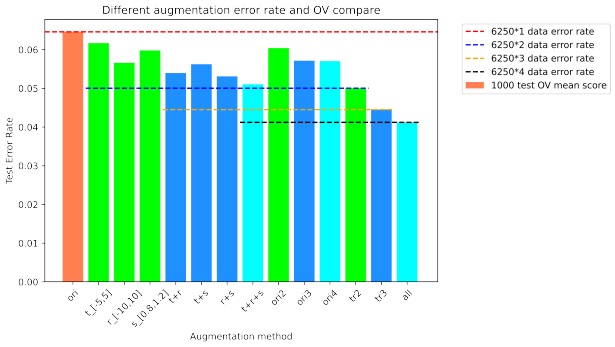
The procedure of the experiment is described as followed. Firstly, the model is trained using Random Forest for all groups. Then we use LIME to get the explanations for the test set. After that, the overlapping(OV) scores are calculated using Equation 2 with $p = 0$, which only takes the percentage of overlapping area over the GT area. The estimated GT is defined as the digit (white/grey) area in images.

It is important to mention that for the group with 1,000 images in test set, overlapping scores are calculated for all the test data. For the group with a test set of 44,000 images, we only count the overlapping score for the corrected ones since a larger test data set could provide us a larger number of corrected images. Otherwise, the OV of limited corrected number will be similar in different training groups.

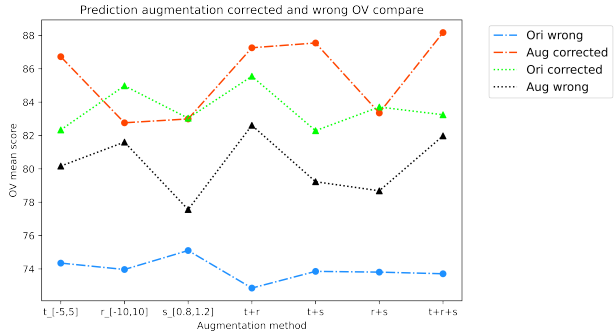
4.1.2. Results and Analysis

The results of this experiment are showed in Figure 3 and Figure 4. There are several finding from those figures. Firstly,

after data augmentation in Figure 3(a), the error rate seems to be decreasing. Secondly, it could be seen that the OV mean score in the corrected ones after augmentation is much higher than the wrong ones using the original data as training, and OV in the wrong ones after augmentation is lower than the corrected ones in the original from Figure 3(b), which meets our expectations. Thirdly, the fitted lines in Figure 4 show the negative correlations between error rate and OV score. Even though the p_{value} for the result of test set with 44,000 images is around 0.56, the fitted line for the test set with 1,000 images seems much more convincing with a $p_{\text{value}} = 3.1 \cdot 10^{-4}$.



(a) The bar chart represents test error rates of the different experiments in the 14 groups. The color represents data size. The horizontal lines are the test error rates of the different control test groups.



(b) OV mean score comparing between different wrongly predicted labels.

Fig. 3. Results for Experiment 1 with OV score and error rate.

The results of this experiment provide strong evidence that data augmentation will reduce the error rate while increasing the OV. The subsequent experiments will continue to explore more complex images.

4.2. Experiment 2: Dogs v.s. Cats Classification

4.2.1. Dataset

In order to extend our experiment to more complex images, we use The Oxford-IIIT Pet Dataset[11] to conduct a dogs v.s. cats classification task. The original dataset has 25 breeds for dogs and 12 breeds for cats as labels, we simplify the

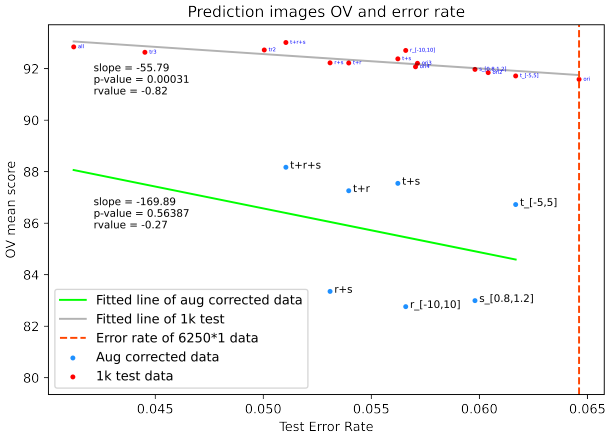


Fig. 4. The fitted line for two kinds of test rest. Both fitted line have a negative slope, the 1k test fitted line $p_{\text{value}} = 0.00031 < 0.05$

problem by combining all dogs and cats, converting it to a binary classification problem. To balance the number of data for each class as well as reduce the computational cost for experiments, sampling has been implemented. Thus, the dataset used in our experiments is a sampled one with 1000 images for each class in training set and 400 images for each class in validation set.

One of the advantages for this dataset is, it also provides the pixel level foreground-background segmentation which could be considered as the ground truth for explanation.

4.2.2. Implementation details

As mentioned in Section 3, here we set several experiments with different kinds of data augmentation methods as well as their combinations. The augmentation operations used in this example includes geometric and photometric methods, namely, (a) rotation (notated as r), within the range of [-90,90]; (b) transformation (notated as t), includes randomly width shifting, height shifting and flipping operations; (c) brightness (notated as b), ranged from 0.5 to 1.5 times of brightness change; (d) channel shift (notated as c), transformations over RGB channels with range 50. All augmentation methods used in this experiment are implemented by `ImageDataGenerator` class provided by `tensorflow`.

The neural network used in this example is a simple structured CNN, whose layers information is provided in Appendix B. We set the optimizer as Adam [12], learning rate as 10^{-5}. Every experiment is set to run 45 epochs with early stop mechanism to avoid over-fitting. p in Equation 2 is set as 0.5 in this case.

4.2.3. *Results and Analysis*

Figure 5 shows the result of this experiment. With the x-axis representing the accuracy and y-axis representing the overlapping score, we can see that they seem to be linearly correlated. The Pearson correlation coefficient is around 0.89 with p-value of 0.0012, presenting that the accuracy and overlapping score is positively correlated with high certainty. The transparency of the cross mark reveals to what degree do these runs implement data augmentation, and the augmented operations are notated beside the cross. We can see a darker cluster in the top-right corner in comparison with the bottom-left one indicating a trend in data augmentation intensity similarly to the overlapping score. Since a higher overlapping score could indicate the generalizing ability of the model, the results tell us that data augmentation aids the performance by improving the understanding (explanation) of the task.

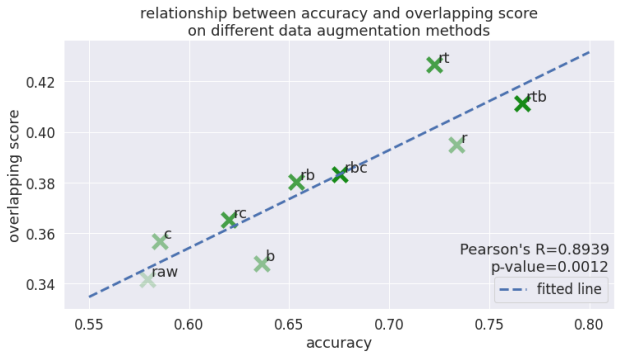
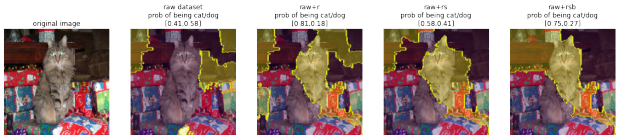


Fig. 5. Overlapping score and accuracy have a high positive correlation in the Pets Sampled dataset. The notations beside cross markers present the augmented operation, e.g. $'rtb'$ means the training dataset includes the raw training set, the rotated set, the transformed set and the set with the change of brightness.

However, the remaining question is, through which aspect does the data augmentation help model explanation? Is it related to the augmented operations? We explore here some breeds with the most corrected labeled images after data augmentation to try to give a qualitative illustration.

The first thing that worth mentioned is that the model with data augmentation seems to perform better when predicting images with complicated lighting. As shown in Figure 6(a), the model with data augmentation is able to better capture the hair illuminated by the sunlight. A second finding is, that with data augmentation, models seem to be separating better the object from the background. In a case like 6(b) where the background of the image is complicated, the model with raw data only focuses on unimportant background parts while after data augmentation the positive area for predicting includes the cat's body.

(a) Shiba Inu case



(b) Maine Coon case

Fig. 6. Case study for Shiba Inu and Maine Coon. The first column shows the original image, while the rest of columns show the positive area (yellow area) for its true label in explanation through LIME. The probabilities of predicting the image as a cat or as a dog is present above each picture.

5. CONCLUSIONS AND FUTURE WORK

Both experiments show a correlation between data augmentation, higher accuracy and higher OV score. We can even see that the slopes in the accuracy/OV plots for both experiments are in the same order of magnitude. Even though this is the case, in order to confirm our hypothesis further, experiments on more complicated datasets and more complicated models are required.

In Experiment 2 we have seen an improvement on OV and accuracy by using data augmentation in our CNN but, due to the simplicity of the problem, we were not able to generalize it to Transfer Learning. When using Transfer Learning the accuracy of the models (InceptionV2 [13], MobilenetV2 [14], Resnet50 [15]) was close to 99% for our dataset leading to no room for improvement with data augmentation.

This paper has provided a new method on how to evaluate image classifier explanation quantitatively using LIME. This new method evaluates the accuracy and the explanation performance of Machine Learning models before and after using Image Augmentation for MNIST and Cats and Dogs classification datasets. The result shows a clear correlation between using Image Augmentation, improving accuracy and improving the explainability (OV) score of the models. The increased OV score could be seen as an explanation on why image augmentation allows the models to generalize better and reduce overfitting. This approach may be generalized to other models and experiments to explain why data augmentation reduces overfitting and improves generalization of the models. The methodology in this paper can also help to evaluate the explainability performance of different image classifiers which can be important with legislations such as GDPR [16].

6. REFERENCES

[1] Luis Perez and Jason Wang, "The effectiveness of data augmentation in image classification using deep learning," *arXiv preprint arXiv:1712.04621*, 2017.

[2] Connor Shorten and Taghi M Khoshgoftaar, "A survey on image data augmentation for deep learning," *Journal of Big Data*, vol. 6, no. 1, pp. 1–48, 2019.

[3] Andrew Selbst and Julia Powles, ""meaningful information" and the right to explanation," in *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, Sorelle A. Friedler and Christo Wilson, Eds., New York, NY, USA, 23–24 Feb 2018, vol. 81 of *Proceedings of Machine Learning Research*, pp. 48–48, PMLR.

[4] Hanjie Chen and Yangfeng Ji, "Improving the explainability of neural sentiment classifiers via data augmentation," *arXiv preprint arXiv:1909.04225*, 2019.

[5] Li Deng, "The mnist database of handwritten digit images for machine learning research [best of the web]," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 141–142, 2012.

[6] Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and C. V. Jawahar, "Cats and dogs," in *2012 IEEE Conference on Computer Vision and Pattern Recognition*, 2012, pp. 3498–3505.

[7] 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*, 2016, pp. 1135–1144.

[8] Alexander B. Jung, Kentaro Wada, Jon Crall, Satoshi Tanaka, Jake Graving, Christoph Reinders, Sarthak Yadav, Joy Banerjee, Gábor Vecsei, Adam Kraft, Zheng Rui, Jirka Borovec, Christian Vallentin, Semen Zhydenko, Kilian Pfeiffer, Ben Cook, Ismael Fernández, François-Michel De Rainville, Chi-Hung Weng, Abner Ayala-Acevedo, Raphael Meudec, Matias Laporte, et al., "imgaug," `https://github.com/aleju/imgaug`, 2020, Online; accessed 01-Feb-2020.

[9] Hao Zhang, Jiayi Chen, Haotian Xue, and Quanshi Zhang, "Towards a unified evaluation of explanation methods without ground truth," *arXiv preprint arXiv:1911.09017*, 2019.

[10] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[11] Omkar M. Parkhi, Andrea Vedaldi, Andrew Zisserman, and C. V. Jawahar, "Cats and dogs," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2012.

[12] Diederik P Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[13] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna, "Rethinking the inception architecture for computer vision," 2015.

[14] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," 2019.

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," 2015.

[16] Maja Brkan, "Do algorithms rule the world? algorithmic decision-making and data protection in the framework of the gdpr and beyond," *International journal of law and information technology*, vol. 27, no. 2, pp. 91–121, 2019.

Appendices

A. DETAILED INFORMATION FOR EVERY RUN IN EXPERIMENT 1

No	Training group	Data size	Description	No	Training group	Data size	Description
1	ori	6250	Training 01 data	8	t+r+s	25000	No 1+2+3+4(aug part)
2	t_[-5,5]	12500	No 1 + translation aug by range [-5,5] rando m at x and y direction	9	ori2	12500	No 1*2
3	r_[-10,10]	12500	No 1 + rotate aug by range [-10,10] random	10	ori3	18750	No 1*3
4	s_[0.8,1.2]	12500	No 1 + scale aug by range [0.8,1.2] random	11	ori4	25000	No 1*4
5	t+r	18750	No 1+2+3(aug part)	12	tr2	12500	Traning 01+02 data
6	t+s	18750	No 1+2+4(aug part)	13	tr3	18750	Traning 01+02+03 data
7	r+s	18750	No 1+3+4(aug part)	14	all	25000	Traning 01+02+03+04 data

Fig. 7. Details about every run in Experiment 1.

B. CNN STRUCTURE USED FOR EXPERIMENT 2

layer name	output size
Input	(150,150,3)
Conv2D_1	(148,148,16)
MaxPool2D_1	(74,74,16)
Conv2D_2	(72,72,32)
MaxPool2D_2	(36,36,32)
Conv2D_3	(34,34,64)
MaxPool2D_3	(17,17,64)
Flatten	(18496)
Dense_1	(512)
Dense_2	(2)

Table 1. CNN structure used for Experiment 2