Explaining the Effect of Data Augmentation on Image Classification Tasks

Jerry Tang
Stanford University
jhtang21@stanford.edu

Manasi Sharma
Stanford University
manasis@stanford.edu

Ruohan Zhang
Stanford University
zharu@stanford.edu

Abstract

Explainability remains a substantial challenge in deep learning research. Data augmentation is a proven tool to improve image classification results, yet very little research has been conducted on its explainability. We applied five data augmentation techniques on 100 MNIST examples and used them to train CNNs along with the baseline model. Saliency mapping is retrieved from each model using 100 MNIST testing images as query. Classification evaluation indicates that several augmentation methods improved accuracy compared to the baseline. Pearson's Correlation Coefficient (CC) and Kullback-Leibler Divergence (KL) were computed to compare information gain/loss of saliency mappings with respect to the baseline model, and they agree on distortion diverging the most from the baseline saliency mappings, but CC and KL no correlation with validation and testing accuracy. Lastly, Shannon entropy is computed for each saliency mapping but no significant variation is found across different models. The conclusion is that data augmentation can result in better models and significantly alter saliency mappings, and that observing saliency mappings qualitative and quantitatively provides intuition on how data augmentation alters models. However, there remains concrete evidence to explore on how saliency mapping can be used to explain and optimize data augmentation. Future research should consider using a similar methodology but with more complex images and radical augmentation methods, as well as including measurements such as RUC and IG.

1. Introduction

Computer vision and artificial intelligence has became a popular field of research following the discovery of deep learning applications on computer vision [19] and the popularization of benchmark platforms and datasets such as ImageNet [6]. One of the major applications of deep learning is the task of identifying objects in a given image, called image classification. Impressively, using ImageNet as a performance benchmark, state of the art neural networks are able to achieve top-5 error rates below 5% [7,22]. While image classification using deep learning models has advanced sufficient enough for widespread real world applications, one remaining challenge is explainability. In this paper, we will apply concepts such as data augmentation an saliency mapping to provide an analysis and potential application to further improve image classification tasks through data augmentation.

1.1. Literature Review

Explainability in artificial intelligence refers to the logical reasoning, understanding, and intuition of what exactly makes highly effective models arrive at their prediction [14]. Deep learning models are difficult for humans to find intuition because of their nested, non-linear structures, and even advanced models are often trained and applied in a black-box nature. This lack of transparency becomes an legal and ethical problem in many high impact applications: for example, when concerning with medical, legal, or safety decisions [5]. Other than gate-keeping the safety and ethics of computer decisions, explainability is also powerful in that it can be used to improve deep learning models by detecting flaws, verifying predictions, and act as way of gaining insights to the problem.

Current solutions for explainability focuses on decomposing model and input variables to measure the relevance of each. The method of sensitivity analysis measures the sensitivity of individual pixels to the correct prediction and can potentially suggest which pixels to alter to improve the correct classification score [18]. However, one weakness of this method is very result oriented and offers little intuitive insight into explainability [10]. The method of Layer-wise relevance propagation is a more complex decomposition than sensitivity analysis that identifies pixels that are the most pivotal to making the correct classification [2].

Saliency methods is a class of tools that highlight the relevant features of an input, such as regions of pixels of a given image that is classified [1]. Saliency maps of images can be produced through a back propagation using already trained convolutional neural networks, by finding the classification weights and taking the absolute value across all channels [18]. Saliency map can also be useful for object segmentation/localization (identifying pixels that belongs to an object in a given image) [18], because in common cases, the saliency map color-highlights the pixels of the classified object. This is done by separating pixels by a certain saliency threshold [3].

Data augmentation is a technique that enhances the quantity and quality of the training data for use in deep learning training [17]. In the task of image classification, popular data augmentation techniques include flipping, cropping, rotating, distortion, color distortions, blurring, and many more. Augmented-generated images retains their original label and is used as additional training data. Data augmentation targets issues that comes along a training dataset that is too small, which leading to overfitting [13].

1.2. Problem Statement

To recall key points; First, explainability remains an important challenge. Second, saliency mapping is a potential tool for studying explainability. Third, data augmentation is a powerful tool in deep learning training to reduce overfitting and enhance models.

However, studies on the explainability aspect of data augmentation is non-existent, especially using saliency mapping. Thus, this study targets using saliency mapping to understand the explainability of data augmentation, with the ultimate goal of optimizing data augmentation and improving image classification accuracy. We will deploy combinations of augmentation techniques to generate various sets of training data based on the MNIST dataset of handwritten digits [9]. An identical convolutional neural network(CNN) model will be trained independently on each augmented dataset with the original images as well as only on the original images. We will then apply saliency mapping on the trained CNN's and conduct quantitative and qualitative analysis. Further details of the methodology will be provided in next section.

This study will result in an analysis to contribute to the Stanford Vision Learning Lab for our future study on supervised data augmentation using saliency mapping.

2. Methodology

2.1. Dataset and Data Augmentation

All training data are directly from or augmented from the MNIST dataset imported through the PyTorch library, which consists of 60,000 training images and 10,000 testing images, with images being a handwritten digit from 0 to 9. Due to the simplicity of MNIST images, models can easily achieve accuracy over 0.99, which leaves little room for experiment and improvement. Thus, we intentionally limit our training data to a fixed-subset of 100 examples, including 10 examples for each digit. Each image retains their original dimension of $1 \times 28 \times 28$.

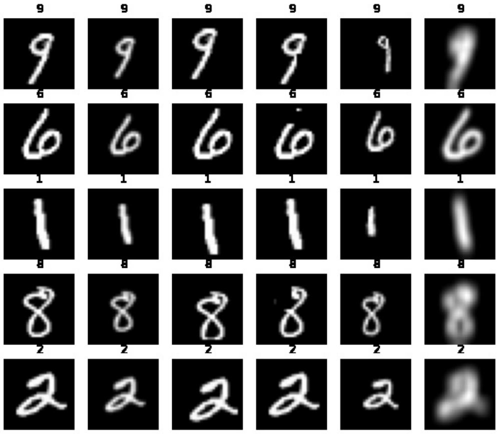


Figure 1. Random examples of augmented images from MNIST. Columns from left to right: original, scaled, randomly cropped, randomly erased, distorted, and blurred. Notice that padding is included to ensure image dimensions are constant.

Five data augmentation methods are applied to the MNIST subset. All data augmentation techniques directly uses PyTorch's built in transformation methods. The first method is random cropping, which randomly crops the image into size $1 \times 24 \times 24$. The second method is Gaussian blurring, which blurs the entire image based on a Gaussian distribution. The third method is random scaling, which we randomly scale the image to widths between [10, 28]. The fourth method is distortion, which we use the RandomPerspective() method to distort the image. The fifth method is random erasing, which random pixels are erased. Fig. 1 illustrates collection of the five methods applied independently on the same randomly selected examples. In all cases, padding is automatically applied to retain image dimensions.

Random data augmentation is done 10 times for each of the five method, generating 1000 examples total. Then, the original 100 examples are added back to each dataset, making a total of 1100 examples for each set of training data, each consisting of 110 examples for each digit. Lastly, the original dataset is concatenated 11 times with itself, making a baseline dataset of size 1100 with only duplicates of un-augmented images. Training is then done on these 6 datasets. The model trained with the original examples is the baseline mode and will be used as benchmark.

2.2. CNN Architecture and Training

We choose a simple CNN architecture that can achieve 0.99 accuracy on the full MNIST dataset within a few epochs. The CNN consists of two convolution layers, each consisting of a convolution layer followed by ReLU activation function and a max pooling layer. The two convolution layers are followed by a Linear and softmax layer.

Identical training procedures are followed to train 6 CNN's based on our architecture. We choose Adam as our optimizer, use learning rate = 0.01, and train the models for 30 epochs. During every step (10 total) in each epoch, the loss is recorded. During every epoch, a validation test is done and recorded. The model with the highest validation accuracy is be saved as the best model. At the end of training, the best model is tested on the testing dataset. The training returns the best model, recorded losses, validation accuracies, best validation accuracy, and the testing accuracy.

2.3. Saliency Map

Gradient-weighted Class Activation Mapping (Grad-CAM) is used as the saliency method [15]. Grad-CAM highlights important regions in an image for classifying as a concept (such as the nose of a dog to classify as a dog), and is directly applicable to our trained CNNs. Grad-CAM will be applied to each of our trained CNNs with query images for evaluations. We have our bespoke Python code following the method in Grad-CAM to apply to our CNN and cater to specific needs such as overlapping the query image and heat map.

2.4. Evaluation

CNN: augmented and original datasets will be trained for 30 epochs, which is sufficient for convergence. Then, respective CNNs will be evaluated based on their training loss, validation results, and testing accuracy.

Grad-CAM: 100 MNIST images from the testing dataset will be used as query images to generate saliency mappings from each trained CNN, yielding a total of 600 saliency maps. Then, quantitative metrics, mainly from [4], and qualitative observations will be retrieved and analyzed for these saliency maps.

Specifically, under the assumption that there is little to no bias in our evenly distributed subset of MNIST, the best metric to pay attention to are Pearson's Correlation Coefficient (CC), which measures correlation and dependency between images. Individual CC is computed for the same query image and between saliency maps of the baseline model and each of the augmented models, where we use the baseline model's saliency mappings as the ground truth fixation map. For saliency image P and ground truth Q, the CC is computed from Eq. (1) [8]:

$$CC(P,Q) = \frac{cov(P,Q)}{\sigma(P) \times \sigma(Q)} \tag{1}$$

Thus, since there are 100 query images and 5 models trained on different augmented datasets, 100 CCs are computed for each model. Despite CNN not being probabilistic, we will also compute the Kullback-Leibler Divergence (KL) score for each CNN with respect to the baseline model, which measures the loss of information compared to the baseline as ground truth. KL is computed with the same pairs of query images as CC, yielding 100 for each model. With ϵ as a regularization constant, KL is computed from Eq. (2) [4]:

$$KL(P,Q) = \Sigma_i Q_i log(\epsilon + \frac{Q}{\epsilon + P}) \tag{2}$$

We then look at the statistics of the CC and KL scores for each model, such as their mean, interquartile range, and standard deviation. Normalized Scanpath Saliency (NSS), Area under ROC curve (AUC), Information Gain (IG), Similarity (SIM), and Earth's Moving Distance (EMD) are either computationally expensive and difficult to optimize or too time-consuming to generate for this project, and thus will not be computed.

In addition to measurements taken from [4], the Shannon entropy will be computed for each saliency map, yielding a total of 600 scores. For saliency map X and pixels $x \in X$, Shannon entropy is computed by Eq. (3):

$$H(X) = -\Sigma_{x \in X} p(x) log(p(x)) \tag{3}$$

We will then look at statistics differences in Shannon entropy between saliency maps trained by 6 CNNs to analysis the effect of data augmentation on Shannon entropy. Similar to CC and KL, we also analyze the mean, interquartile range, and standard deviation regarding each model. Shannon entropy is a measurement of the uncertainty and randomness of an image [16], and because is a probabilistic based measurement, we will normalize image pixels to sum up to 1 before computing their Shannon entropy.

3. Results and Discussion

This section is broken up to discuss the data augmentation results, saliency mapping results, and quantitative evaluation. Augmentation resulted accuracy and classification results are presented in the first subsection, augmentation's influence on saliency mapping is visualized in the second subsection, and the last two sections present quantitative measurements as indicated above.

3.1. Data Augmentation Training Results

Certain data augmentation techniques improved model performance compared to the un-augmented baseline. The training loss is illustrated in Fig. 2. We can see that most of the CNNs converge close to zero within 15 epochs. Out of which, the original training dataset converged the fastest, due to only composing of 100 unique images. The outlier is the dataset with perspective transformation (distortion),

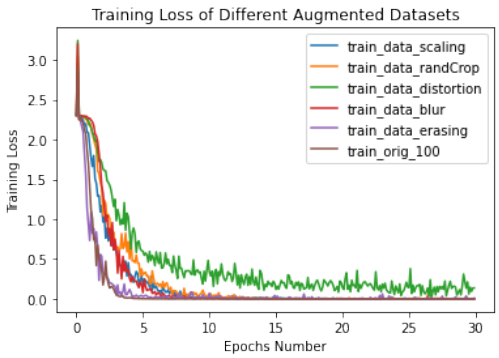


Figure 2. Loss on the training data recorded while training. Unmodified data converges the fastest while augmented datasets are slow to converge or doesn't converge closer to zero.

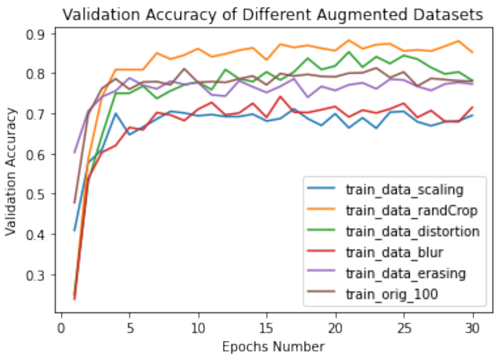


Figure 3. Validation accuracy converge to different levels. Random cropping is consistently higher than the other datasets. The original dataset performs around average.

Dataset	BestVal	Test
Scaling	0.71	0.67
RandCrop	0.88	0.87
Distortion	0.85	0.79
Blur	0.74	0.73
Erasing	0.79	0.79
Orig	0.81	0.78

Table 1. Training Results

which converges to a higher loss at a slower pace. This is likely because while perspective transformation is useful in object detection to mimic different camera angles [12, 20], the camera angle in the rest of the experimenting datasets is fixed, except for the perspective transformation dataset, bringing it a unique training challenge.

Validation accuracy is recorded every epochs and is plotted in Fig. 3. The convergence speed is approximately equal across all datasets, but converged accuracy varies. Random cropping consistently leads and achieves close to 0.90 accuracy. Despite converging to higher loss, the distortion dataset performs the second best, suggesting that its model is generalized. The blurring and scaling datasets perform worst due to unexpected noisy data. The blurring dataset performs poorly likely due too much blur causing digits to be mixed, considering their shapes are similar to begin with, and the scaling datasets performance is likely due to the fact that the validation set only contains images of the full size and lacks images of different scales.

The best validation accuracy and testing accuracy is logged in Tab. 1, and we see that performance on testing is similar of validation. The augmentation experiment suggests that mild augmentation such as cropping and perspective translation improves model performance, while excessive augmentation, such as likely the case of blur, can add too much noise. This is especially true for a simple dataset like MNIST.

3.2. Saliency Maps and Qualitative Results

We have 100 query images to generate 100 corresponding saliency mappings for each CNN. We took 5 samples query images' saliency mapping and illustrated in Fig. 4. Each column represents 5 samples generated using the same CNN, and different between columns is the training data used to train the CNNs. From the left most to right most column, the training data is: original, scaled, randomly cropped, randomly erased, distorted, and blurred. When visualizing saliency mappings, it is helpful to see where the attention is concentrated with respect to the original query images. Therefore, we also generated an overlapped image for each saliency map by combining with their query image, as illustrated in Fig. 4a. On the other hand, only the saliency map is needed as queries for quantitative analysis, and therefore samples such as ones in Fig. 4b is preserved for evaluations below.

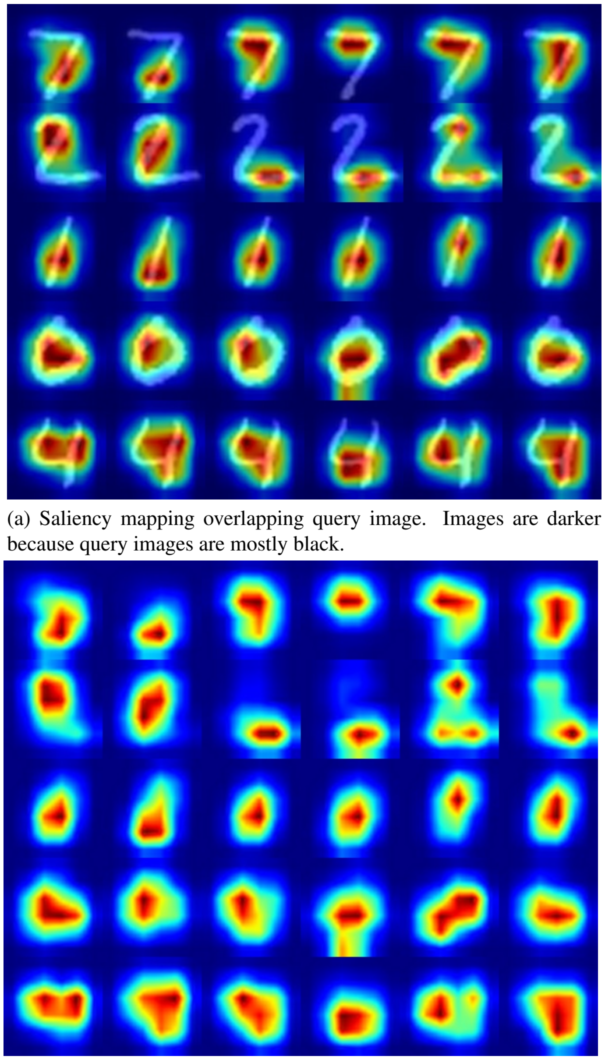
Saliency results in the generated samples demonstrates several generalized qualitative results. First, attention focus regions are strictly on or next to the handwriting, while the background is largely unattractive. This indicates that the models have almost no background bias. Second, the attention hot spots does not cover the entire handwriting, which is expected because training data handwriting varied in size and position, with or without data augmentation. Third, when comparing each of columns 2-6 (augmented trained models) with the first column (baseline model), difference in saliency mapping between models clearly exists and is only possibly contributed by the difference in augmented training data. This indicates that data augmentation directly changes the attention of model, and with at least a few techniques that enhances testing accuracy (Tab. 1), it reasonably supports the argument that shifting to optimize attention increases model accuracy and is an explanation to why data augmentation is useful.

Note that, based on the examples we generated, it is difficult to make a qualitative judgement to accurately summarize and reason the specific differences. Specific qualitative judgements for every saliency image, such as human annotations of regions or subjective ratings, is not practical due to the large number of images and the fact that this project only have one primary author. Nonetheless, saliency mapping visualizations and qualitative assessments is helpful in providing an intuition to support the explainability aspect of deep learning [11], and concrete evidence is supported by the quantitative results in the following section.

3.3. CC and KL Evaluation

CC is computed for each saliency mapping from model trained on augmented data compared to the baseline. Fig. 5a graphs the distribution for each model in boxplots, and Tab. 2 highlights the mean and standard deviation for CCs across the same model. In this application, CC treats saliency maps as variables and measures the correlation and dependency between the two, and high CC score occurs when values from the augmented saliency map is close to the baseline saliency map in magnitude [4]. In our case, we can treat our models as random variables, with 100 CCs being samples. Moreover, since CC is measured for each model against the baseline model, the average CC should tell us how correlated and dependent an augmented model is to the baseline model in terms of their saliency mappings.

We see that scaling and erasing as augmentation techniques does preserves saliency CC the most, with means above 0.800. This is likely because scaling just about completely preserves all features of the original image, which allows the CNN to ”look for” the same features. One exception where scaling can significantly alters features is if very few pixels remained, which can be a problem for our dataset since the images were only 28 times 28 to begin with. However, scaling is minimized at 10 times 10 pixels by design to preserve digits’ integrity. On the other hand, erasing preserved high CC likely because we picked our eraser size (0.05, 0.10) to be quite small in order to prevent augmented images to be unreasonably ineligible. Random cropping and blurring results in only slightly lower CC and much of the four models thus far fall under a similar range based on interquartile range and standard deviation Fig. 5a. On the other hand, distortion results in a significantly lower mean CC score as well as the largest standard deviation. This shows that image distortion as data augmentation, even maintaining eligibility from the human eye to maintain the correct classification, will cause significant changes to the attention of the model. Interestingly, even with a lower CC, the model with distortion is tested to be slightly more accurate than the baseline model in terms of both validation and testing accuracy (Tab. 1). Lastly, the larger standard deviation of the distortion model remains as another point of interest. One possible explanation is that distortion introduces more randomness to general image features, which is not only reflected in the mean CC but also a larger standard deviation within a model’s CC scores.



(a) Saliency mapping overlapping query image. Images are darker because query images are mostly black.

(b) Saliency mappings with heat map only. Similar samples are used as query for quantitative evaluation.

Figure 4. Visualize saliency mapping results through 5 samples for each of 6 CNNs we trained. Images within a column is generated from the same CNN trained by same datasets. Training dataset from left to right (column): original, scaled, randomly cropped, randomly erased, distorted, and blurred.

KL is computed with same query pairs as CC, where one saliency map is from an augmented model and the other from our baseline model as ground truth. The distribution of KL scores for each model is graphed in Fig. 5b. Since KL measures the loss of information in respect to the baseline saliency maps, a lower KL score would resemble a

Model/Dataset	CC Mean	CC STDV
Scaling	0.802	0.100
RandCrop	0.768	0.126
Distortion	0.620	0.154
Blur	0.737	0.119
Erasing	0.828	0.105

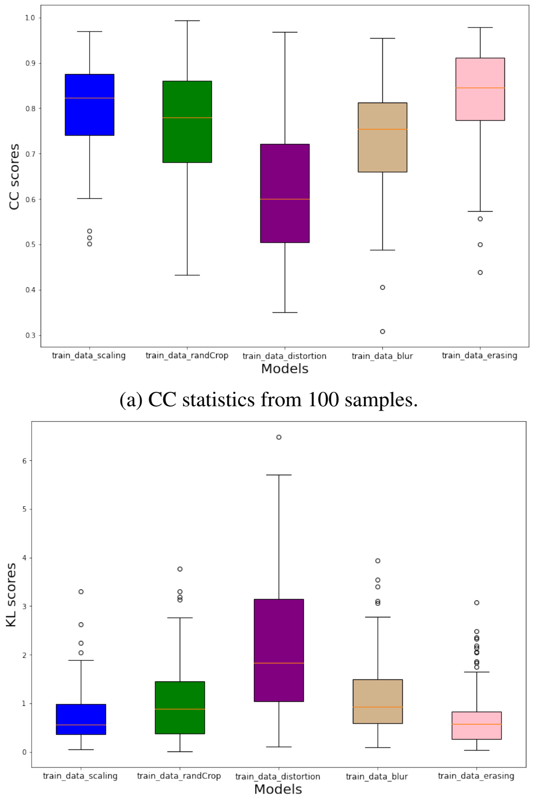
Table 2. Pearson's Correlation Coefficient (CC) mean and standard deviation for each model.

Model/Dataset	KL Mean	KL STDV
Scaling	0.736	0.557
RandCrop	1.050	0.822
Distortion	2.163	1.436
Blur	1.151	0.781
Erasing	0.729	0.636

Table 3. Kullback-Leibler Divergence (KL) mean and standard deviation for each model.

better approximation to the ground truth by the augmented saliency map. We see from Tab. 3 that the order of data augmentation techniques in terms of mean KL score of their saliency samples, from highest to lowest is: distortion, blurring, random cropping, scaling, and erasing. Interestingly, this is the exact opposite order for their CC scores. This is reasonable, as KL measures the loss of information, while CC measures the preservation of information, and they are intuitively opposites. Distortion also resulted in the largest KL standard deviation, but is even larger in proportion to the other standard deviations compared to CC. The conclusion for analyzing KL scores is similar to that of CC: which is that we see clear influence of data augmentation altering the saliency mappings.

The agreement in results between the CC and KL scores is somewhat surprising, as CC is a linear statistical method while KL is a probabilistic statistical method. With regard to the question of whether a higher CC/KL is more optimal, the answer is that, based on mean CC/KL scores and model validation and testing accuracy, there is no obvious correlation between the two and therefore no preference. The reason for this non-correlation is that the ground truth saliency mapping should not be used as the correct saliency mapping to preserve, and diverging from the baseline saliency mapping can result in either better or worse classification accuracies. If we could obtain a correlation between Shannon entropy and classification accuracy, it would mean that data augmentation can be optimized by supervising its change in Shannon entropy.



(a) CC statistics from 100 samples.

(b) KL statistics from 100 samples.

Figure 5. Pearson's Correlation Coefficient (CC) and Kullback-Leibler Divergence (KL) for saliency maps from models trained with augmented datasets compared with saliency maps of the same query image from the baseline model.

Model/Dataset	SE Mean	SE STDV
Scaling	4.989	0.283
RandCrop	5.082	0.284
Distortion	4.964	0.291
Blur	5.063	0.259
Erasing	5.102	0.281
Orig100	5.125	0.281

Table 4. Shannon entropy(SE) statistics for each model.

3.4. Shannon Entropy Evaluation

Shannon entropy is computed for every saliency map generated by our models, including the baseline model. The distribution of Shannon entropy is graphed in Fig. 6 and key statistics are in Tab. 4. Shannon entropy should explain the randomness of saliency mappings under different models [21], but the results suggests that there is no significant difference in Shannon entropy across models.

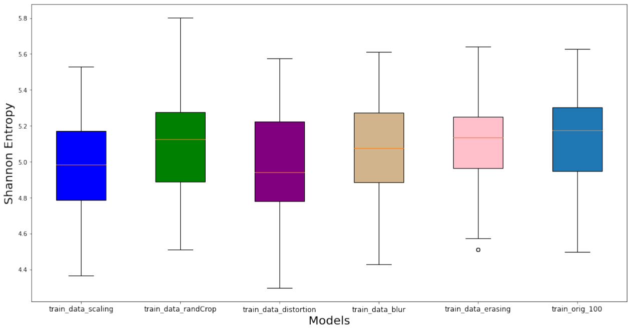


Figure 6. Distribution of Shannon entropy for each model indicates that difference is insignificant.

4. Conclusion

Data augmentation is an emerging tool in deep learning research, and its explainability remains a substantial challenge in research. For this study, we use saliency mapping as an entry point, with statistical analysis, as an attempt to explain the effects of different data augmentation techniques on image classification. We used MNIST as our dataset, with only 100 samples used as training data. Then, five data augmentation techniques were applied: scaling, random cropping, distortion, blurring, and random erasing. Six CNNs with identical architecture are trained with augmented datasets as well as the original set. Saliency mappings are retrieved with each model using 100 query images from the testing dataset.

Initial data augmentation results after 30 epochs suggests training with augmented datasets converges slower, likely due to more sophistication in the dataset. This might present a greater challenge for more sophisticated image datasets, which is that data augmentation will further increase training time [17]. As for classification results, random cropping consistently achieves higher validation and testing accuracy, while a few other techniques have also shown to be potentially useful. Our hypothetical reasoning for this is that due to the fact that MNIST is very simple and the baseline model already performs well, therefore excess augmentation only adds noise to the data.

Saliency mappings provides a intuitive visualization of the effect of data augmentation on model attention, where we see that all saliency maps have little bias and is very focused on regions of the handwriting. There is, however, obvious shift in heat map between models trained with different datasets. To quantify the effect of augmented models, CC and KL scores were computed, and it was found that distortion augmentation method most heavily alters the information contained in saliency maps compared to the baseline and creates the most randomness. However, we see no correlation between CC and KL scores with classification accuracy. To independently quantify saliency maps from each model, Shannon entropy is computed on normalized saliency maps to approximate the level of information in each saliency map, but there is no significant difference across models.

To conclude our findings, data augmentation still requires more explanation. On one hand, it is easier to understand the effect of data augmentation of saliency mappings, since we know the exact augmentation techniques and parameters being applied. On the opposite hand, saliency mapping comes from the black box that is the CNN, which makes it harder to explain data augmentation in reverse. The use of CC/KL and Shannon entropy tells us that different data augmentation techniques alters saliency mappings to different degrees, which in term impact classification decisions. However, it is still unclear on how saliency mapping can be used as a tool to control and optimize data augmentation.

We propose future studies to continue to look into the use of saliency mapping to explain data augmentation. First, we suggest increasing overall difficulty for the model by both using a more sophisticated image set and also using more radical data augmentation techniques. One issue with this specific study is the use of MNIST limits the potential for data augmentation since the images are very simple and the baseline is too high. Second, more information can be obtained from saliency maps by computing different measures, especially RUC and IG. Third, mass human annotation can be conducted to evaluate qualitative results based on subjective metrics. Lastly, different models can be tested to compare the impact of data augmentation on different architectures.

5. Appendices

All code, images, and results can be found and reproduced from my GitHub page.

6. Contributions & Acknowledgements

J.T. conducted all experiments, wrote all codes, wrote all text, made all figures and graphs, and conducted all analysis work. J.T., M.S., and R.Z. are all affiliated with the Stanford Vision Learning Lab and this project is part of a greater research on data augmentation with the guidance of saliency mappings. R.Z. advised the project and assigned the general project idea, and J.T. implemented them with modifications. M.S. provided example codes for retrieving saliency mappings and computing measures that were modified by J.T. for this project's needs. M.S. also provided guidance and advisory.

References

[1] Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity checks for

saliency maps. *Advances in neural information processing systems*, 31, 2018. 1

[2] Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one*, 10(7):e0130140, 2015. 1

[3] Yuri Y Boykov and M-P Jolly. Interactive graph cuts for optimal boundary & region segmentation of objects in nd images. In *Proceedings eighth IEEE international conference on computer vision. ICCV 2001*, volume 1, pages 105–112. IEEE, 2001. 2

[4] Zoya Bylinskii, Tilke Judd, Aude Oliva, Antonio Torralba, and Frédo Durand. What do different evaluation metrics tell us about saliency models? *IEEE transactions on pattern analysis and machine intelligence*, 41(3):740–757, 2018. 3, 5

[5] Jaegul Choo and Shixia Liu. Visual analytics for explainable deep learning. *IEEE computer graphics and applications*, 38(4):84–92, 2018. 1

[6] 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. 1

[7] 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. 1

[8] Olivier Le Meur, Patrick Le Callet, and Dominique Barba. Predicting visual fixations on video based on low-level visual features. *Vision research*, 47(19):2483–2498, 2007. 3

[9] Y LeCun, C Cortes, and C Burges. The mnist dataset of handwritten digits (images). *NYU: New York, NY, USA*, 1999. 2

[10] Grégoire Montavon, Wojciech Samek, and Klaus-Robert Müller. Methods for interpreting and understanding deep neural networks. *Digital Signal Processing*, 73:1–15, 2018. 1

[11] T Nathan Mundhenk, Barry Y Chen, and Gerald Friedland. Efficient saliency maps for explainable ai. *arXiv preprint arXiv:1911.11293*, 2019. 5

[12] Jesús Munoz-Bulnes, Carlos Fernandez, Ignacio Parra, David Fernández-Llorca, and Miguel A Sotelo. Deep fully convolutional networks with random data augmentation for enhanced generalization in road detection. In *2017 IEEE 20th International Conference on Intelligent Transportation Systems (ITSC)*, pages 366–371. IEEE, 2017. 4

[13] Luis Perez and Jason Wang. The effectiveness of data augmentation in image classification using deep learning. *arXiv preprint arXiv:1712.04621*, 2017. 2

[14] Wojciech Samek, Thomas Wiegand, and Klaus-Robert Müller. Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models. *arXiv preprint arXiv:1708.08296*, 2017. 1

[15] 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. 3

[16] Claude Elwood Shannon. A mathematical theory of communication. *ACM SIGMOBILE mobile computing and communications review*, 5(1):3–55, 2001. 3

[17] Connor Shorten and Taghi M Khoshgoftaar. A survey on image data augmentation for deep learning. *Journal of big data*, 6(1):1–48, 2019. 2, 7

[18] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013. 1, 2

[19] Athanasios Voulodimos, Nikolaos Doulamis, Anastasios Doulamis, and Eftychios Protopapadakis. Deep learning for computer vision: A brief review. *Computational intelligence and neuroscience*, 2018, 2018. 1

[20] Ke Wang, Bin Fang, Jiye Qian, Su Yang, Xin Zhou, and Jie Zhou. Perspective transformation data augmentation for object detection. *IEEE Access*, 8:4935–4943, 2019. 4

[21] Yue Wu, Yicong Zhou, George Saveriades, Sos Agaian, Joseph P Noonan, and Premkumar Natarajan. Local shannon entropy measure with statistical tests for image randomness. *Information Sciences*, 222:323–342, 2013. 6

[22] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1492–1500, 2017. 1