Investigation of the Effects of Medical Image Augmentation on the Robustness of Convolutional Neural Networks

A summary of Undergraduate Dissertation in Computer Science with Artificial Intelligence

Jakub Adrian Niemiec
BSc (Hons) Computer Science with
Artificial Intelligence
Department of Computer Science
Brunel University London

***Abstract*—In this report we investigate the effects of image augmentation on a LeNet-inspired Convolutional Neural Network (CNN) model to understand how augmentation of skin nevus images affects the accuracy of binary classification.**

Augmented images are fed into the network model to measure how it responds to varying levels of image augmentation in terms of final classification accuracy and loss. The results highlight the benefit of utilising rotation as the preferred method for augmenting nevus image datasets; resulting in improved accuracy and reduction of misclassification of malignant samples.

We also demonstrate how the model focuses on specific elements of an image and show how noise could be used as an adversarial form of attack on a network forcing it into misclassification - highlighting the adverse effects on utilising black-box AI agents in medicine.

Keywords*—*artificial intelligence, ai, machine learning, overfitting, adversarial attacks, image augmentation, convolutional neural networks, vision, computer vision, medicine

I. Introduction

Novel technologies led to the emergence of medical solutions utilising Artificial Intelligence; detecting diseases and assisting doctors in diagnosing diseases. A known example being Google's DeepMind state-of-the-art models analysing eye scans and predicting eye diseases - a proof-of-concept demonstrating how machine learning can improve early disease detection rates and raise the standard of healthcare [1].

Cancer has been a major health concern for many years; it's estimated that, in 2018 alone, there has been over 1.7 million new cases of cancer-related diagnoses in the United States with thousands more worldwide, therefore, improving detection rates can translate to improved longevity and quality of healthcare globally [2]. However, research indicates that Deep Neural Networks (DNNs) can be manipulated into misclassification after minimal image augmentation [3]. This stipulates that image classification models are prone to errors and thus are not as reliable as we first thought; a seemingly insignificant overfitting error can lead to a malignant tumour being misdiagnosed as a benign. If such advanced computer implementations are to be used in healthcare systems, it calls for an investigation of the effects that overfitting can have on the utility of artificial agents.

Skin cancer is one of the most common types of cancer in the world [4] diagnosed using visual inspection by a trained dermatologist. If the human expert believes the nevus to be malignant the patient is prescribed a dermoscopic analysis or biopsy [5].

In a study investigating the classification of several sub-types of skin cancer, a CNN model achieved $72.1 \pm 0.9\%$ accuracy, compared to two human dermatologists who attained 65.56% and 66.0% respectively thus encouraging the use of artificial agents in aiding diagnosis [5].

However, DNNs used for image classification can suffer from overfitting, resulting in the network becoming prone to misclassifications if the model obsesses over sample characteristics irrelevant to its goal [6]. Such unpredictability of a black-box DNNs is a cause for concern when considering that machine learning is used for disease detection.

It is necessary to understand the fundamental issue with model overfitting to better understand the inner workings of an, otherwise, black-box model used for medical diagnosis. To do so, it is necessary to develop an experiment whereby a network is 'attacked' by a corrupted dataset, investigating how it affects the model's ability to classify correctly. This investigation exposes a trained model to such "attacks" to understand how they can benefit or harm the classifier.

Overfitting has been investigated under various conditions to expose the tendency of CNNs to fixate on irrelevant properties of the dataset, causing poor generalisation translating to sub-par performance on previously unseen data. It becomes apparent that, though minimal changes to the dataset, a CNN can be "tricked" into misclassification; such findings must be treated as a red flag, considering that CNN-based systems are used in healthcare - network models used for assistive medical diagnosis are prone to the same issues. These findings warrant further investigation into the nature of misclassification by CNNs to find ways for those black-box models to generalises well enough to minimise the vulnerability to adversarial attack.

II. Method

A. Neural Network Design

The model utilised in this experiment had to satisfy the complexity parameters to ensure that it was capable of processing images and extracting properties necessary for generalisation of their features - a step necessary to make predictions on future, unseen, samples.

In this investigation the CNN model was based on the LeNet [7] architecture; originally used for classification of handwritten characters and a known example of the utility of CNNs in optical character recognition. The original architecture was adjusted and its feature space changed to increase the classification accuracy on unaugmented dataset samples which would be used as a comparison "base" model. At its core the network consists of an input layer, three hidden layers and a SoftMax classifier, a simple albeit effective structure primarily dictated by the computational resources available.

B. Experiment Design

The experiment consists of a series of augmentations being performed on the original dataset each with varying intensity to demonstrate how varying levels of augmentation affect the CNN's ability to classify the samples.

Each augmentation level consists of four sub-experiment runs to expose the variation of behaviour of the network to augmentation in training, testing and both dataset sub-sets. The table below outlines how the datasets are prepared prior to experiment run.

	Corruption/Augmentation Enabled?	
	Training Data	*Test Data*
Control	No	No
Training Corrupted	Yes	No
Validation Corrupted	No	Yes
All Corrupted	Yes	Yes

TABLE I. Dataset Augmentation Application Plan for Each Experiment Type

The CNN model is exposed to each of the four sub-experiments with varying (controlled) degree of augmentation applied to the relevant sub-set of the samples, the resulting accuracy and loss metrics on the unseen (validation) data are then compared against the base model which was only exposed to the non-augmented samples. This comparison demonstrates how augmentation affects the CNNs ability to generalise the samples and make correct classifications.

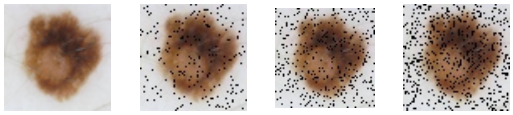
C. Image Corruption

The first portion of the experiment investigates how image corruption, through artificial noise, affects the CNNs ability to make correct classifications. The dataset is manually augmented prior to being fed into the model; all images are rescaled to the network's input requirements and noise in the form of black pixels is added at random co-ordinates across the image in accordance with the corruption factor. The table below outlines the minimum, maximum and step increase of corruption factor across the dataset samples.

	Medical Dataset
No. of Samples	2,160
Min. Level of Corruption	0.0 (0%)
Max. Level of Corruption	0.15 (15%)
Step Increase	0.05 (5%)

TABLE II. Image Corruption Application Properties

Applied augmentation results in black pixels being applied to the samples prior to exposing them to the CNN, examples of how this form of corruption appears on a medical dataset can be seen below.



0% 5% 10% 15%

TABLE III. REPRESENTATION OF IMAGE CORRUPTION AT VARIOUS LEVELS WITH EXTREME SAMPLES FOR MEDICAL DATASET

D. *Image Rotation*

The image rotation experiment has a similar structure to the pixel corruption. However, when performing the rotations, the dataset is not pre-prepared as in the case of the noise augmentation, instead a built-in Keras library function pre-processes the images on-the-fly as they are fed into the network model.

The image pre-processing library loads the image and rotates it randomly within a given range in either positive or negative direction depending on a random integer. For example, given a setting of "45" the image could be rotated in a range of -45 ° to 45 ° depending on the random integer generated when the image is loaded by the program.

This experiment utilises a the entire range of 180 °, allowing for a full 360 ° random rotation of the image at 45° increments - the image below shows examples of rotated nevus images from the dataset.

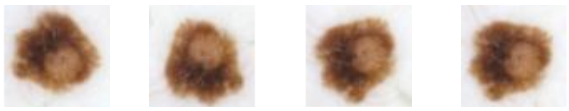


TABLE IV. REPRESENTATION OF RANDOM ROTATION SAMPLES OF THE SAME IMAGE

III. RESULTS

Utilising rotation augmentation results in a reduction of error rate by 15% against no augmentation at all, meaning that the model performs better when rotation is applied across the dataset. The True Positive (TP) rate shows that using rotation results in TP rate of 77% as compared to only 43% and 62% for noise augmentation and no augmentation respectively. Suggesting that augmenting the dataset using rotation allows the model to generalise more effectively and thus become more robust at classifying skin cancer correctly.

Introduction of noise to the dataset increased the likelihood of misclassification; the base error rate raised by 3% making the network more likely to misclassify samples, however using rotation decreased the error rate by 2%.

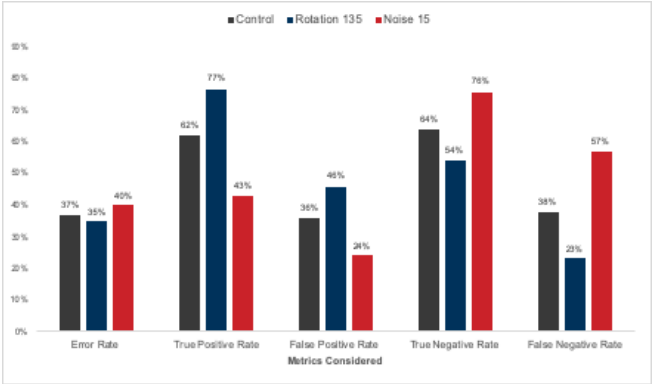


Fig. 1. A Summary Bar Graph Summarising Comnfusion Matrix Calculations for Experiments Yielding Highest Accuracy

The results suggest that using rotation on the dataset can results in a model capable of better classification and lower error rates as compared to using no augmentation at all or using noise. Conversely, it demonstrates that networks are vulnerable to image distortion in the form of noise which can lead to misclassification and, by extension, misdiagnosis.

The grid below shows samples from the original dataset and their respective heat maps demonstrating the areas the network "focuses" on whilst learning the features of the dataset.

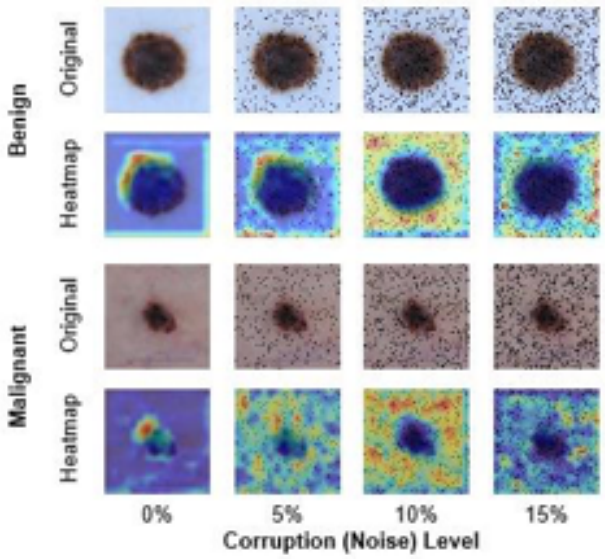


Fig. 2. A Grid Showing Activation Heatmap Overlay for Varying Types of Image Augmentation – Noise Only

The final network accuracy scores show an overall drop in accuracy of the classifications made by the model; this corresponds to the heat map visualisation as demonstrated by the dissipation of hot regions away from the object of classification to the surrounding area indicating the model learning features of the area around the nevus rather than the properties of the nevus itself.

This behaviour correlates with accuracy metric for the model at each noise augmentation level - with highest accuracy 0% noise (control) and lowest accuracy of 57.59% at 10% augmentation.

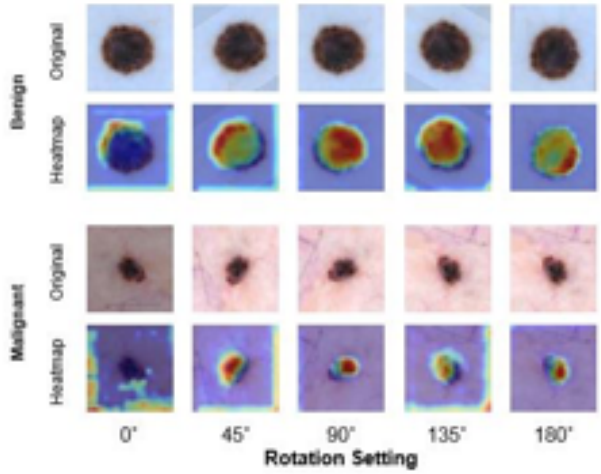


Fig. 3. A Grid Showing Activation Heatmap Overlay for Varying Types of Image Augmentation – Rotation Only

Heat maps for rotational augmentation show a contrasting model behaviour to that of noise augmentation; the model appears to focus better on the nevus object whilst ignoring the background features. Based on the experimental results, the rotation setting of 135 ° achieved the highest accuracy of 65.35% with some loss around the edges of the image (lighter blue areas).

The horizontal lines visible in the benign sample at 0% is likely attributed to the measurement scale embedded into the endoscopic device used to take the photographs, it is an example of how the model can lean towards static elements in the background - straying away from the objective of the classification task. As image is rotated the embedded measurement scale is likely to be cut out or distorted thus forcing the model to discard it as a feature of interest.

The use of heat maps helps to visually justify the statistical metric results collected from the experiments, they also highlight ways in which adversarial attack can be draw away the model's "focus" from the classification object by introducing augmentation.

IV. Discussion

The comparison of augmentation methods with varying intensities for a binary classification task demonstrates how the process of augmenting dataset samples can affect the model's accuracy and loss. Moreover, it further points out potential exploits in machine learning models which can be used as an adversarial attack vector to induce misclassification. This section focuses on evaluation of augmentation methods which scored the highest binary classification accuracy in their respective experiments and their comparison to aid deduction of conclusions with regards to the augmentation method yielding best results.

	Val. Acc.	**Val. Loss (BCE)**	**Error Rate**	**Type II Error Rate**	**True Positive**
No Augmentation	63%	63.95%	37%	38%	62%
Rotation (135°)	65.35%	63.79%	35%	23%	77%
Noise (15%)	59.81%	65.91%	40%	57%	42%

TABLE V. Summary Table of Confusion Matrice Calculations based on Experimental Results

Based on the summary table above, it is evident that the use of rotation can have beneficial effect on the overall accuracy of the model. Accuracy of classification increased by 2.35% during validation stages with the true positive rate for this type of dataset augmentation showing a 15% increase in correct classification of malignant cases which turned out to be cancerous. Significantly, Type II Error Rate score for rotation augmentation techniques showed a decrease by 15%, implying that the network is less likely to diagnose a malignant sample as benign which, given the medical application of the model, would result in potentially terminal disease progressing further and endangering the patient who's diagnosis has been missed.

Conversely, the results suggest that noise augmentation decreases classification accuracy (dropping from control of 63% down to 59.81%) meaning that noise can be used to force the model to misclassify the samples as part of an adversarial attack vector. Noise caused the model to struggle, increasing its error rate by 3% with a Type II Error Rate increasing by 19% - these results, combined with the binary cross-entropy loss metric, demonstrate the augmentation of dataset images with black pixels can decrease the robustness of a machine learning model.

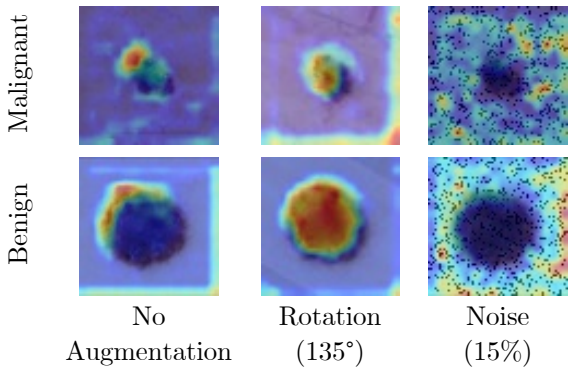


TABLE VI. HEATMAP COMPARISON FOR EXPERIMENTS YIELDING HIGHEST ACCURACY OF CLASSIFICATION

This conclusion is further supported by activation heat maps extracted from the network. As seen in the table above, rotating images causes the network to focus on the central element of the sample which contains the mole rather than "obsess" over the surrounding area – the likely reason for the network not "paying attention" to the skin around the mole is that, due to rotation, the background changes more frequently as compared to the mole itself (which is always expected to be in centre frame).

Samples with noise have introduced additional elements to the image which the model focussed on instead of generalising the nevi at the centre. The heat maps suggest that the model obsessed over the noise rather than the nevi sample which is evident by the way in which focus hotspots dissipate from the mole itself to the area where noise was introduced.

In conclusion, the comparison of both dataset augmentation methods clearly suggests that utilising sample rotation can increase the efficiency of the model whilst decreasing the error rate when compared to the control dataset with no augmentation. Contrastingly, the introduction of black pixel noise to the dataset samples causes the model to be more likely to misclassify the samples, in the case of medical datasets this means that the model can classify cancerous nevi as benign which will have adverse impact on the health of the patients involved.

REFERENCES

[1] De Fauw, J. et al., 2018. Clinically applicable deep learning for diagnosis and referral in retinal disease. p. 41.

[2] National Cancer Institute, 2018. *Understanding Cancer - Cancer Statistics.* [Online] Available at: https://www.cancer.gov/about-cancer/understanding/statistics [Accessed October 2018].

[3] Nguyen, A., Yosinski, J. & Clune, J., 2015. Deep Neural Networks are Easily Fooled: High Confidence Predictions for Unrecognizable Images. *IEEE.*

[4] American Cancer Society, 2016. *Cancer Facts & Figures 2016,* s.l.: American Cancer Society Inc..

[5] Esteva, A. et al., 2017. Dermatologist-level classification of skin cancer with deep neural networks. *Nature,* 02 February, Issue 542, pp. 115-118.

[6] Szegedy, C. et al., 2014. Intriguing properties of neural networks. *ArXiv Preprint,* pp. 1-10.

[7] LeCun, Y., Boyyou, L., Bengio, Y. & Haffner, P., 1998. Gradient-Based Learning Applied to Deocument Recognition. *Proceedings of the IEEE.*

[8] Brownlee, J., 2016. *Machine Learning Mastery: What is a Confusion Matrix in Machine Learning.* [Online] Available at: https://machinelearningmastery.com/confusion-matrix-machine-learning/ [Accessed 14 03 2019].

[9] Brownlee, J., 2016. *Overfitting and Underfitting With Machine Learning Algorithms.* [Online] Available at: https://machinelearningmastery.com/overfitting-and-underfitting-with- machine-learning-algorithms/

[10] [Accessed 04 01 2018].

[11] Caruana, R., Lawrence, S. & Giles, L. C., 2000. Overfitting in Neural Nets: Backpropagation, Conjugate Gradient, and Early Stopping. *Advances in neural information processing systems,* Issue 13, pp. 402-408.

[12] de Boer, P.-T., Kroese, D., Mannor, S. & Rubinstein, R. Y., 2005. A Tutorial on the Cross- Entropy Method. *Annals of operations research,* 134(1), pp. 19-67.

[13] DeepAI, n.d. *Neural Network.* [Online] Available at: https://deepai.org/machine-learning-glossary-and-terms/neural-network [Accessed 28 12 2018].

[14] Ghotra, M. S., Dua, R. & Pentreath, N., 2017. Mean Squared Error and Root Mean Squared Error. In: *Machine Learning with Spark - Second Edition.* s.l.:Packt Publishing.

[15] Goodfellow, I. J. et al., 2014. Generative Adversarial Nets. Advances in Neural Information Processing Systems 27 (NIPS 2014), pp. 1-9.

[16] Google, 2018. *Machine Learning - Classification: Accuracy.* [Online] Available at: https://developers.google.com/machine-learning/crash- course/classification/accuracy [Accessed 26 01 2019].

[17] Google, n.d. *A history of machine learning.* [Online] Available at: https://cloud.withgoogle.com/build/data-analytics/explore-history-machine- learning/#footnote_23 [Accessed 29 12 2018].

[18] Huang, S. et al., 2017. Adversarial Attacks on Neural Network Policies. *arXiv:1702.02284.* ISDIS, 2018. *ISIC Archive.* [Online] Available at: http://display.isic-archive.com/#!/onlyHeaderTop/gallery

[19] ISDIS, 2018. *ISIC Project.* [Online] Available at: https://isdis.net/isic-project

[20] JetBrains, 2019. *TeamCity.* [Online] Available at: https://www.jetbrains.com/teamcity/ [Accessed January 2019].

[21] Kaggle, 2013. *Kaggle Dogs vs. Cats.* [Online] Available at: https://www.kaggle.com/c/dogs-vs-cats [Accessed 2018].

[22] Karpathy, A., 2018. *Convolutional Neural Networks for Visual Recognition.* [Online] Available at: https://cs231n.github.io/convolutional-networks/ [Accessed 29 12 2018].

[23] Karpathy, A., 2018. CS231n Convolutional Neural Networks for Visual Recognition. [Online]|Accessed 25 01 2019].

[24] Keras, 2019. *Keras: The Python Deep Learning Library.* [Online] Available at: https://keras.io/[Accessed Februrary 2019].

[25] Keras, 2019. *Usage of metrics.* [Online] Available at: https://keras.io/metrics/ [Accessed 14 03 2019].

[26] Krizhevsky, A., Sutskever, I. & Hilton, G. E., 2012. ImageNet Classification with Deep Convolutional Neural Networks. *Advances in neural information processing systems,* 25(2).

[27] Kuhn, M. & Johnson, K., 2013. In: *Applied Predictive Modeling.* s.l.:Springer; 1st ed. 2013, Corr. 2nd printing 2018 edition, p. 256.

[28] Kumar, A., 2018. *QA: Blackbox Testing for Machine Learning Models.* [Online] Available at: https://dzone.com/articles/qa-blackbox-testing-for-machine-learning-models [Accessed February 2019].

[29] Kurakin, A., Goodfellow, I. J. & Bengio, S., 2016. Adversarial Examples In The Physical World. *arXiv:1607.02533,* Issue 4.

[30] Mirza, M. & Osindero, S., 2014. Conditional Generative Adversarial Nets. *arXiv preprint arXiv:1411.1784.*

[31] Moosavi-Dezfooli, S.-M., Fawzi, A. & Frossard, P., 2016. DeepFool: A Simple and Accurate Method to Fool Deep Neural Networks. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR),* Volume 6, pp. 12103-12117.

[32] Negnevitsky, M., 2011. Artificial Intelligence: A Guide to Intelligent Systems. Third Edition ed. s.l.:Addison Wesley.

[33] OpenCV, 2019. *OpenCV - About.* [Online] Available at: https://opencv.org/about.html [Accessed January 2019].

[34] Panchal, F. S. & Panchal, M., 2014. Review on Methods of Selecting Number of Hidden Nodes in Artificial Neural Network. *International Journal of Computer Science and Mobile Computing,* November, 3(11), pp. 455-464.

[35] Plunkett, K. & Elman, J. L., 1997. Exercises in Rethinking Innateness: A Handbook for Connectionist Simulations. In: s.l.:MIT Press, p. 166.

[36] Rastegari, M., Ordonez, V., Redmon, J. & Farhadi, A., 2016. XNOR-Net: ImageNet Classification Using Binary Convolutional Neural Networks. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics),* Volume 9908 LNCS, pp. 525-542.

[37] Rosenblatt, F., 1958. The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review,* 65(6), pp. 386-408.

[38] Selvaraju, R. R. et al., 2017. Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization. *IEEE International Conference on Computer Vision (ICCV).*

[39] Skymind, n.d. *A Beginner's Guide to Neural Networks and Deep Learning.* [Online] Available at: https://skymind.ai/wiki/neural-network#concept [Accessed 29 12 2018].

[40] Su, J., Vargas, D. V. & Sakurai, K., 2017. One Pixel Attack for Fooling Deep Neural Networks. *arXiv:1710.08864.*

[41] TensorFlow, 2018. *TensorFlow Guide.* [Online] Available at: https://www.tensorflow.org/guide [Accessed 2018].

[42] TensorFlow, n.d. *Get Started with TensorFlow.* [Online] Available at: https://www.tensorflow.org/tutorials/ [Accessed 2018].

[43] The International Society for Digital Imaging of the Skin, 2018. *ISIC Project.* [Online] Available at: https://isdis.net/isic-project [Accessed 09 2018].

[44] van Gerven, M. & Bohte, S., 2018. Artificial Neural Networks as Models of Neural Information Processing. s.l.:Lausanne: Frontiers Media.

[45] Wang, J. & Perez, L., 2017. The Effectiveness of Data Augmentation in Image Classification using Deep Learning.