

Study on Data Augmentation Method for Brain Tumour Detection

Yuxuan Li *

University of Wisconsin Madison, 502 N Frances St, Madison, Wisconsin, USA

* Corresponding author: li2234@wisc.edu

Abstract. Brain tumours are very dangerous diseases in the world. Brain tumours will cause people to lose their brain function and the survival rate of having brain tumours is low. The characteristics of brain tumours are very complex, so, the accuracy of manual classification to classify benign and malignant brain tumours comes with low accuracy. In this case, the classification method with the use of machine learning (ML) becomes significant since ML classification gets higher accuracy than manual classification. Among ML models, Convolutional Neural Networks (CNN) are widely used to classify brain tumour pictures. In ML, strong models have more parameters and more data are required to train those parameters. However, datasets may not be big enough to train the model. Therefore, few-shot learning methods are proposed to increase the accuracy of the model. In this study, three different methods of data augmentation are used to augment 3,000 pictures' brain tumour datasets. The codes are successfully run, and the graph of accuracy is exhibited. In this article, the result of the study shows that the cutout method implemented in data augmentation performs the best in detecting brain tumours.

Keywords: Data Augmentation, Brain Tumour, Machine Learning, Tumour Detection.

1. Introduction

Brain tumours are diseases that harm children's and adults' health. Brain tumours make up about 2 percent of total cancer deaths in the U.S. and those tumours are the most common solid tumour for children [1]. Brain tumours refer to a collection of abnormal cells in the human brain. Human skulls are rigid. Thus, the growth such as abnormal cells will be dangerous and life-threatening. Brain tumours are classified as cancerous ones and noncancerous. Uncancerous brain tumours refer to brain tumours that grow slowly. Although uncancerous brain tumours could cause some problems to humans, they can be removed by a proper surgeon since they have a clearly defined border. The possibility that uncancerous brain tumours come back is also low. However, a cancerous brain tumour is malignant. The malignant brain tumours grow rapidly and can spread to other parts of the human body. Malignant brain tumours are very dangerous to humans and many people die due to brain tumours every year. In this case, the correct classification of brain tumour pictures is very important. With a precise diagnosis, more and more people can receive correct treatment on time. Also, people also do not need to suffer anxiety due to a wrong diagnosis.

The pictures of brain tumours are very complex and full of tiny details. So, the accuracy of manual detection is relatively low compared with Machine Learning (ML) detection. ML with a model trained by a great amount of brain tumour pictures can classify the brain tumours more accurately than a human. In recent years, the Convolutional neural network (CNN) model is widely used in classifying brain tumours and receive excellent results. CNN modes like VGGnet [2] and MobileNet [3] are models which receive excellent results in the image classification field. In one study, the CNN model with trained by a brain tumours dataset and receive the 84.19% validation accuracy at best [4]. Thus, using CNN model to classify brain tumours is proven to have a high accuracy. In machine learning, the stronger model needs more data to train its parameter to perform well. However, in some cases, the dataset provided to the model is not big enough, which will influence the efficiency of the model. Thus, a deep learning method named few-shot learning is proposed to increase the accuracy of the model when the dataset is limited. In this article, data augmentation, which is one of few-shot learning methods, is discussed. Sometimes, a poorly trained model may recognize one picture with different

angles as two different pictures. This happens due to the lack of samples. Data augmentation is used to increase the dataset to better train the model.

This study employed three different data augmentation methods, namely normal augmentation (simply flip up-down and left-right, rotate, scale images to create new images), mix up and cutout on VGGnet, it is indicated that cutout augmentation makes the model have the highest accuracy. So, if people are trying to use one method to augment the brain tumours datasets to train the CNN model, it is suggested that cutout will get higher accuracy.

2. Method

2.1. Method introduction

The dataset in this study contains 3,000 images. In ML, 3,000 images are not good enough for training a model. When training a model in ML, big dataset usually has a better result than small dataset. In order to enlarge the dataset to better train the vgg-16 model, the study uses three different data augmentation skill to increase the dataset from 3,000 to 6,000 and then uses the augmented dataset to train the model to get accuracy and loss. Data augmentation means building algorithms through the introduction of unseen data [5]. This study includes three data augmentation method, which will be explained separately. The first method used in this article is named normal augmentation. It is a simple way in python to achieve data augmentation. The normal augmentation in this article uses rotation, flip, scale, and translation to create new datasets. Shorten says that shifting images left, right, up, and down is a very efficient way to prevent positional bias [6]. The second method used in this study is called mixup. Mixup method refer to select two samples randomly and create the new one. Liang says mixup method has a better ability of generalization than normal data augmentation method [7]. The third method is named cutout, which means randomly covering an input image with a square. Cutout is good at improving CCN models' overall performance and cutout is also a commonly used regularization method [8]. In this case, cutout is a suitable method in this experiment.

2.2. Dataset preparation

The dataset used in the study are from internet [9]. The dataset contains two folders (yes and no), and each folder contains 1500 images of brain tumours, including both tumorous and non-tumorous.

First, the pre-processing is necessary since the size of pictures in this dataset is different. So, the first step in the model is resize all the pictures to $156 * 156$. Then the normalization is done to increase the training speed. In this study, 2400 images are used as train set and the rest 600 images are used as validation set.

The model used in this study is named vgg-16. Vgg-16 is a widely used model in CNN and the architecture of the vgg-16 used in this study is showed up in table 1. Convolutional Neural Network (CNN) is a neural network in deep learning which has one or more layers. Wu says that CNN is efficient in computer vision and ML problems [10]. Convolution neural network is mainly composed of the input layer, convolution layer, activation function, pooling layer, full connection layer, and loss function and the main function of CNN is featuring extraction. Due to this function, in recent years, CNN is widely used in image classification field. Brain tumours' classification is a popular topic in medical field and CNN is proved to be effective in brain tumour classification. The vgg-16 is an object detection and classification model. It is effective in image classification and is used in many fields. Vgg-16 has a great number of parameters and in this study, all the parameters are default parameters without any change. This study uses keras in TensorFlow to build the neural network, pyplot to draw the graphs and os to process the path.

Table 1. The architecture of the employed vgg-16

Layer (type)	Output Shape	Param #
vgg16 (Functional)	(None,4,4,512)	14714688
Flatten	(None,8192)	0
Dense	(None,1)	8193

2.3. Experiment process

After pre-processing and data augmentation, this study uses the augmented to train the vgg-16 model separately with 30 epochs. In every code, including the model trained by unaugment data, showed the train and validation accuracy, train and validation loss on y axis and epochs on x axis. The relationship between accuracy and epochs and the relationship between loss and epochs are clearly shown in the graph. In this study, the results between each augmentation method and the trends of accuracy and loss with epochs are compared to get the final conclusion.

3. Results and Discussion

3.1. Performance for the proposed model

Table 2. Three model's accuracy and loss table after trained with different augmented data after 30 epochs

	Train Accuracy	Validation Accuracy	Train Loss	Validation Loss
Normal Augmentation	0.9942	0.9583	0.0168	0.1585
Mixup	0.9929	0.9775	0.0220	0.0977
Cutout	1.0000	0.9975	1.1697e-04	0.0142

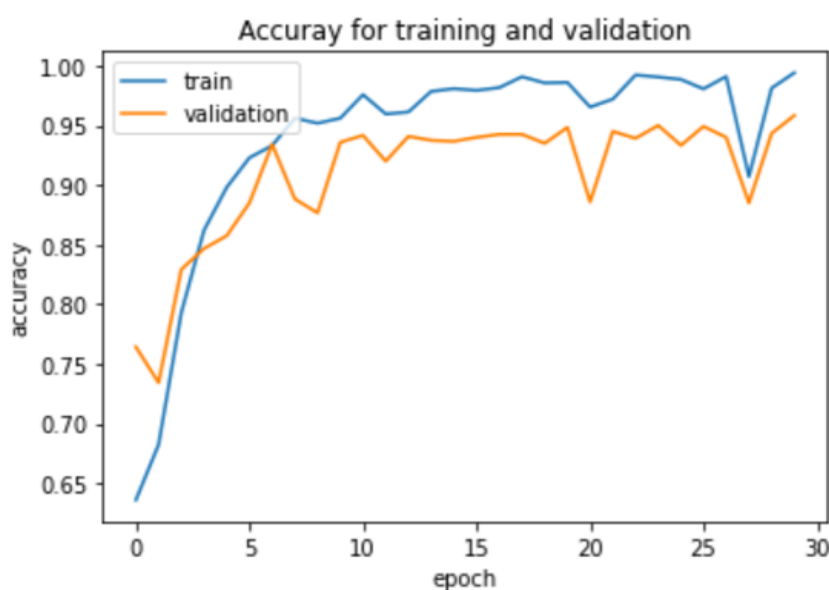


Figure 1. Accuracy for training and validation of model trained by normal augmented data.



Figure 2. Loss for training and validation of model trained by normal augmented data.

Table 2 clearly indicates the accuracy and loss for 3 same models trained by different augmented data using different method after 30 epochs. In figure 1 and figure 2, the accuracy and loss of vgg-16 trained by normal augmented data are showed up. In figure 1, the trend of both accuracy and validation accuracy is increasing. The train accuracy starts from about 0.64 and the validation accuracy starts from about 0.76. After 30 epochs training, the train accuracy comes to 0.9942 and the validation accuracy comes to 0.9583. In figure 2, it is showed that both train loss and validation loss is decreasing when epochs are increasing. At the end of training, train loss reaches to 0.0168 and the validation loss reaches to 0.1585.



Figure 3. Accuracy for training and validation of model trained by mixup augmented data.

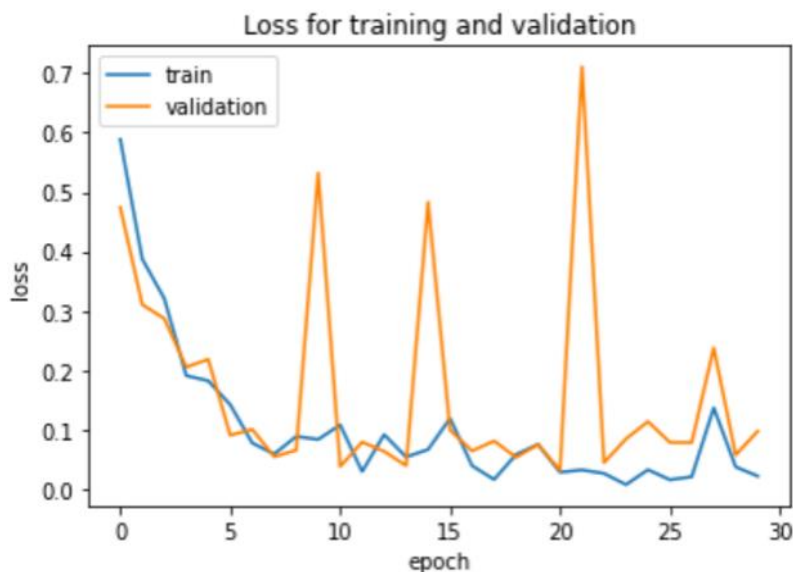


Figure 4. Loss for training and validation of model trained by mixup augmented data.

According to the figures above (Figure 3&4), train accuracy and validation accuracy are both increasing. Train accuracy starts from 0.71 and validation accuracy starts from 0.81. During the increase process, the graph of validation accuracy fluctuates dramatically while the graph of train accuracy increases stably. The final train and validation accuracy go up to 0.9929 and 0.9775. Also, in the loss graph, the validation loss fluctuates dramatically while train loss decreases stably. The final train and validation loss go down to 0.0220 and 0.0977.



Figure 5. Accuracy for training and validation of model trained by mixup augmented data.



Figure 6. Loss for training and validation of model trained by mixup augmented data.

According to the figures above (Figure 6&7), train accuracy and validation accuracy are both increasing. Train accuracy starts from 0.71 and validation accuracy starts from 0.85. After 30 epoch training, train accuracy goes up to 1.000 and validation accuracy reaches 0.9975, which is the highest validation accuracy of the 3 data augmentation method.

By comparing three accuracy graphs and three loss accuracy and table 2, it is indicated that for all models trained by different augmented data, both train and validation accuracy are increasing during training and both train and validation loss are decreasing during training. In this study, for each code's result, all the variables are the same except the augmented method to gurantee that each code's accuracy and loss is affected by the augmented method. The validation accuracy of vgg-16 trained with data without augmentation is 0.9633. The accuracy got by normal augmentation is lower than this accuracy and the accuracy of both mixup and cutout method are higher than this accuracy. Among all three methods, cutout has the highest accuracy. This result may due to the reason that cutout allows the model to consider image context more, such as some tiny features rather than mainly depend on major features. The difference of malignant and benign tumors may exist in small details instead of main features.

4. Conclusions

Nowadays, CNN is employed in many fields because of its powerful image classification function. In medical field, CNN is employed to detect brain tumors. The reason is that manual detections receive low accuracy. This article discusses how different data augmentation method will influence the accuracy of the CNN model when classifying brain tumors. Data augmentation will improve the model's performance and generalization when the dataset to train the model is small. In this study, three different methods of data augmentation are employed to augment the dataset (normal augmentation, cutout augmentation and mixup augmentation.) Among all those methods, models trained by cutout and mixup perform better than the model trained by original dataset, while models with normal augmentation has lower accuracy than model trained by original dataset. And among all three methods, cutout has the greatest accuracy. If someone try to improve the accuracy of CNN model on brain tumor classification field, cutout data augmentation is the most efficient way to help them achieve their goal.

References

- [1] Ostrom, Q., Fahmideh, M., Cote, D., Muskens, I., Schraw, J., Scheurer, M., & Bondy, M. (2019). Risk factors for childhood and adult primary brain tumors. *Neuro-Oncology*, Volume 21, Issue 11, Pages 1357 – 1375. <https://doi.org/10.1093/neuonc/noz123>.
- [2] Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.
- [3] Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., ... & Adam, H. (2017). Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*.
- [4] Abiwinanda, N., Hanif, M., Hesaputra, S.T., Handayani, A., & Mengko, T. (2018). Brain Tumor Classification Using Convolutional Neural Network. *World Congress on Medical Physics and Biomedical Engineering*, Volume 68/1. ISBN: 978 - 981 - 10 - 9034 - 9.
- [5] Dyk, D., & Meng, X. (2012). The Art of Data Augmentation. *Journal of Computational and Graphical Statistics*, 10:1, 1-50. <https://doi.org/10.1198/10618600152418584>.
- [6] Shorten, C. & Khoshgoftaar, T. (2019). A survey on Image Data Augmentation for Deep Learning. *Journal of Big Data* 6,60. <https://doi.org/10.1186/s40537-019-0197-0>.
- [7] Liang, D., Yang, F., Zhang, T., & Yang, P. (2018). Understanding Mixup Training Methods. *IEEE*, Volume 6, pp. 58774 - 58783. <https://doi.org/10.1109/ACCESS.2018.2872698>.
- [8] DeVries, T. & Taylor, G.W. (2017). Improved Regularization of Convolutional Neural Networks with Cutout. *arXiv preprint arXiv: 1708.04552*.
- [9] Whale (2022). Brain tumor detection dataset. <https://www.heywhale.com/mw/dataset/61d3e5682d30dc001701f728/content>.
- [10] Wu, J. (2017). Introduction to convolutional neural networks. *National Key Lab for Novel Software Technology*. Nanjing University. China, 5 (23), 495.