Proceedings of the 2023 International Conference on Software Engineering and Machine Learning
DOI: 10.54254/2755-2721/8/20230197

The study of transfer learning effectiveness on medical brain tumor images

Yizhou Bi

The department of Data Science, Duke Kunshan University, Suzhou, 215300, China

yizhou.bi@dukekunshan.edu.cn

Abstract. The segmentation of brain tumor is an important and challenging issue in the last decade. This study explores the use of transfer learning in medical image segmentation, specifically in the segmentation of lower-grade gliomas in brain Magnetic Resonance Imaging (MRI) scans. The LGG segmentation dataset, sourced from The Cancer Imaging Archive, was utilized for the study. Common data augmentation techniques were applied, such as vertical and horizontal flipping and image scaling, to better simulate real-world scenarios. Transfer learning was used with three widely-used Convolutional Neural Network (CNN) models (i.e. MobileNet, ResNet, and VGG) to establish three groups for each model, with each group consisting of a model that inherits weights and one that does not. The effectiveness of transfer learning was evaluated by training each model until it reaches steady-state, with models that inherit weights undergoing fine-tuning. The Dice coefficient was used as the loss function for the network. The study focused on the accuracy of the model in terms of true positive and true negative. The results showed that transfer learning with weight inheritance generally outperforms transfer learning without weight inheritance, with the VGG model exhibiting the highest accuracy. Overall, the study demonstrates the potential of transfer learning for medical image segmentation tasks, particularly in the context of brain tumor segmentation.

Keywords: transfer learning, convolutional neural network, magnetic resonance imaging, brain tumor.

1. Introduction

A brain tumor is a mass or growth of abnormal cells that develop in the brain. These cells grow uncontrollably and can interfere with normal brain functions such as thinking, sensation, movement, and coordination. Brain tumors can cause a wide range of symptoms, depending on their location, size, and rate of growth. Some common symptoms include headaches, seizures, vision or hearing problems, difficulty speaking or understanding speech, weakness or numbness on one side of the body, and changes in mood or behavior. To diagnose a brain tumor, there are three common methods. (1) Neurological exam, (2) imaging test, (3), biopsy. This study will be focusing on using Convolutional Neural Network (CNN) on the Magnetic Resonance Imaging (MRI) in the imaging test section to diagnose the brain tumor and distinguish it from normal tissues. The convolutional neural network is commonly utilized for the task of identifying abnormal segments in brain images due to its proven performance in semantic segmentation. Compared to previous methods, like linear regression, decision tree, or support vector machine, a convolutional neural network does not rely on human-defined features and generalizes

© 2023 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/8/20230197

features during training by optimizing its filters. However, because labeling MRI for the purpose of neural network training requires intense laboring of highly skilled doctors, the consequential scarcity of the labeled MRI data puts the vanilla supervised learning at a disadvantage [1, 2], potentially limiting the accuracy and depth of the CNN model. Moreover, given the complexity of MRI data or any other medical image, and the size of modern-day neural network, the training may require extensive amount of computational resources to optimize the model in reasonable amount of time. Nowadays, technics like, transfer learning, semi-supervised learning, learning from noisy labels, and learning from computer-generated labels [3], are widely applied in research to relieve the needs for data, and other limitations regarding the usage of CNN. Among all the technics mentioned above, this study focuses on the transfer learning. Transfer learning, in the context of neural network, means that using the knowledge from solving one task to solve another similar task [3], where the similarity of the tasks differs. Compare to the full training process of the building a neural network from ground up, the transfer learning only requires moderate amount of fine-tuning using the training data of the targeted task [4].

Study in the field of medical image analysis shows the great advantage of convolutional neural network compared to previous methods. Thanks to its convolution layer, pooling layer, and activation function design, CNN shows great potential in image recognition when first purposed at 1989 [5]. The model works well in image because the convolution layer iterates through the image and pooling layer find the best match. Recent study also shows the potential of very deep convolution model in image recognition [6]. Because CNN, like many other types of neural networks, does not require hand-crafted features, whereas the algorithm of the same era heavily lean on those, when applied to medical image analysis, those complex features can be learned from data during training [7, 8]. In recent years, CNN is widely studied and experiment with medical and clinical purpose. Because of the difficulties in training and acquiring labeled data, many other technics are applied. Among which, the transfer learning becomes most trendy for its performance [1]. Some studies have tried to address the validity of transfer learning with the knowledge domain shift from general image to medical image, but it does not research the scenario where transfer learning is applied to the ischemic stroke [2, 3, 9]. Despite those research does not show direct evidence of transfer learning is effective in prediction of ischemic stroke, they still suggest a favorable view towards the application of transfer learning to improve model performance.

The objective of the research is to verify if a transferred neural network model that is trained with general purpose dataset, for example, ImageNet, is applicable to a medical purposed one, and to what extent, the technic of transfer learning would help improve the accuracy of the prediction of the model. Despite some assessment has been done on whether the shift in knowledge domain during the transfer learning would harm the model performance or not, none of them directly evaluate the transfer learning in the context of identifying brain tumor.

2. Methodology

2.1. Data collection

The author utilized the LGG segmentation dataset, an established dataset that had been used in various studies in recent years [10, 11]. The dataset comprised brain MR images with corresponding manual FLAIR abnormality segmentation masks, sourced from The Cancer Imaging Archive (TCIA). The images corresponded to 110 patients included in The Cancer Genome Atlas (TCGA) lower-grade glioma collection, with at least fluid-attenuated inversion recovery (FLAIR) sequence and genomic cluster data available. The dataset included tumor genomic clusters and patient data in the data.csv file, while all images were provided in .tif format with three channels per image. For most cases (101 out of 110), three sequences were available, namely pre-contrast, FLAIR, and post-contrast (in this order of channels), while for the remaining cases, the post-contrast sequence was missing (nine cases), or the pre-contrast sequence was missing (six cases). Missing sequences were replaced with FLAIR sequence to ensure that all images were three-channel. The masks were binary, one-channel images that segmented FLAIR abnormality present in the FLAIR sequence, available for all cases. Some example images of the MRI result could be found in Figure 1, Figure 2 and Figure 3.

DOI: 10.54254/2755-2721/8/20230197

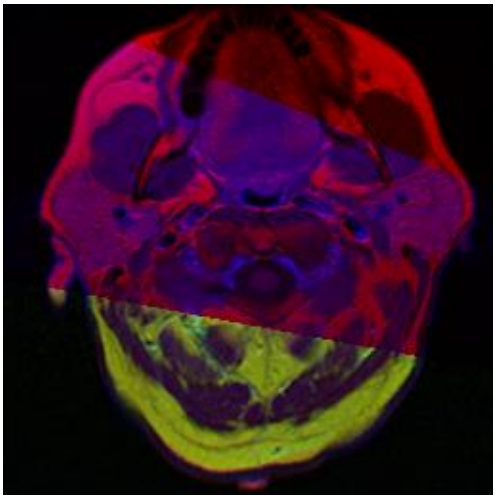


Figure 1. A healthy MRI result with no brain tumor.

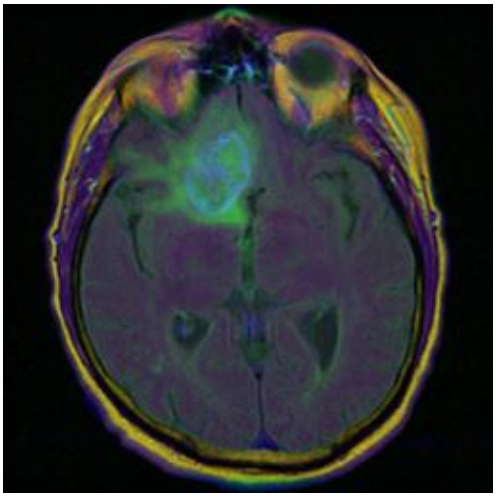


Figure 2. An unhealthy MRI result with brain tumor.

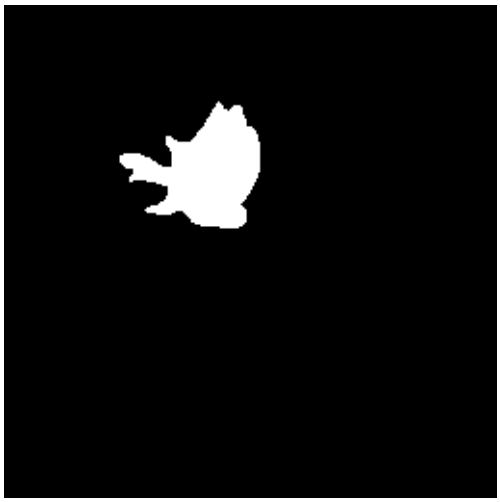


Figure 3. The mask of the unhealthy MRI result with brain tumor.

DOI: 10.54254/2755-2721/8/20230197

2.2. Data augmentation

Common data augmentation techniques were employed in this study. Since MRI images were mostly symmetrical with some orientation variation, the image data were not rotated during training. However, they were vertically and horizontally flipped to better simulate real-world scenarios. In addition, scaling of the images was applied since MRI comprised a set of cross-sectional images of the human brain, and the upper part of the head tended to be smaller, while the lower parts were larger in the imaging. Therefore, exposing the CNN filters to different sizes was advantageous.

This study utilized the built-in preprocessing function of the models being employed. These functions offered pixel value centering and scaling functionality. In the case of the VGG model, the preprocessing function also converted RGB values to BGR values, which was what the model expected. It was worth noting that without these preprocessing steps, the resulting training may not have yielded the expected outcomes, as the input needed to match how the base models were trained.

2.3. Model designing

In transfer learning, the top layers of the base model were usually removed, and the last layer of the remaining base model might not have matched the format of the labeled data. In this study, the output of the VGG, MobileNetV2, and ResNet models was in the shape of (8, 8), while the labeled image had the shape of (256, 256, 1). To transform the former into the latter, the author employed convolutional transpose layers that functioned like the reverse of convolutional layers. These layers increased the feature map dimensions and decreased the filter dimensions, thus allowing the top layers to adapt to the labeled data. Moreover, the author utilized dropout and batch normalization layers to potentially improve training performance and simulate real-world scenarios. Figure 4 presents the model architecture of VGG with customized top layers. In addition, the overview of the model parameters is also shown in Figure 5.

In addition to transposing the base model's last layer to the size of the input, the model also used the skipping connection inspired by U-Net. The skip connections linked the contracting and expansive paths, which enabled the network to leverage both local and global information to make accurate predictions. The contracting path comprised convolutional and max pooling layers that were used to extract features and reduce the spatial dimensions of the input image. Meanwhile, the expansive path consisted of convolutional and up-sampling layers that recovered the spatial resolution of the image and produced a segmentation mask.

DOI: 10.54254/2755-2721/8/20230197

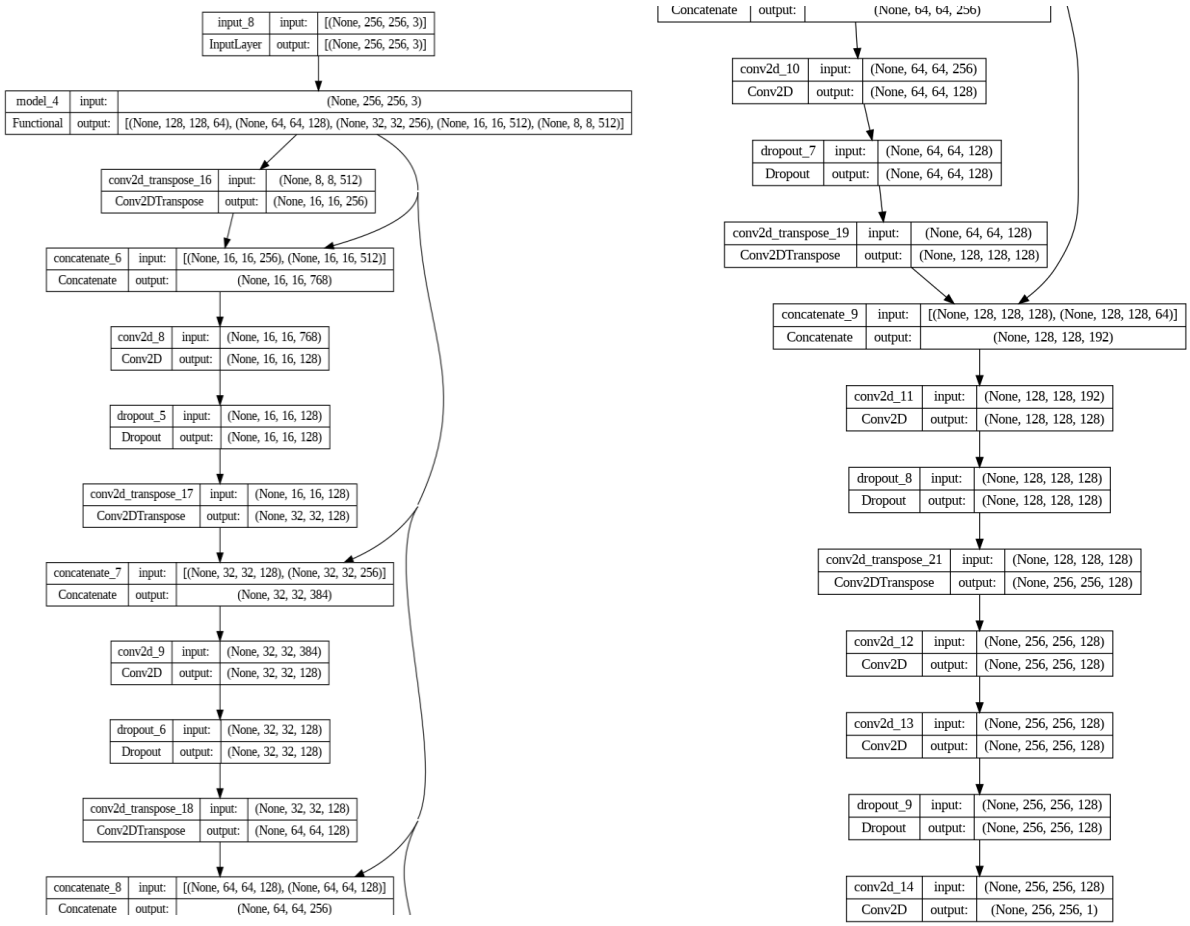


Figure 4. VGG model architecture with customized top layers.

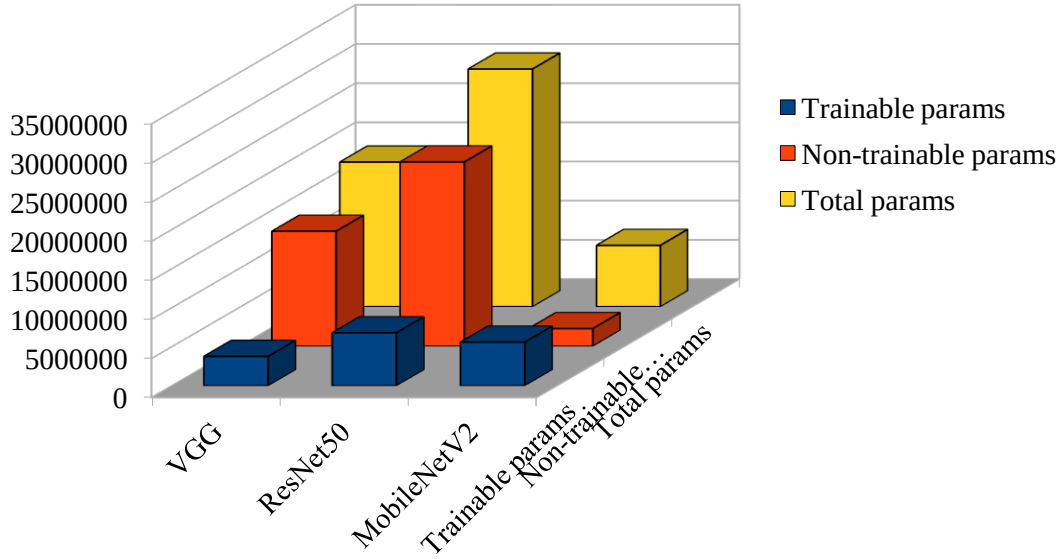


Figure 5. Model (W/ weight) parameter number overview.

DOI: 10.54254/2755-2721/8/20230197

The author had used the Dice coefficient as the loss function for the network. The Dice coefficient was measuring the similarity between two sets of data and was commonly used to evaluate the performance of image segmentation models. It was the ratio of the area of overlap between the predicted segmentation and the ground truth segmentation to the total area of both segmentations. The Dice coefficient ranged from 0 to 1, with 1 indicating perfect overlap between the predicted and ground truth segmentations. Despite its widespread use in segmentation tasks, it was not provided by Tensorflow as a built-in function.

$$Dice(A,B) = 2 \vee A \cap B \vee \frac{}{|A|+B\vee} \tag{1}$$

$$Dice_{loss(A,B)} = 1 - 2 \vee A \cap B \vee \frac{}{|A|+B\vee} \tag{2}$$

2.4. Experiment design

MobileNet, ResNet, and VGG were some of the most widely-used CNN networks for transfer learning. To evaluate the effectiveness of transfer learning, this essay established three groups for each model, with each group consisting of a model that inherited weights and one that did not. Each model was trained until it reached steady-state, which was defined as the point where the model's loss no longer decreased. Models that inherited weights then underwent fine-tuning, where the base model was unfrozen and retrained until it reached steady-state again. This process helped the training to better simulate real-world applications.

In this experiment, the study was interested in the accuracy of the model in terms of true positive and true negative because of its application in the medical field, where accuracy was crucial. If the transfer learning model without weight inheritance performed better than its inherited counterpart, it would conclude that weight inheritance was not particularly helpful, and vice versa.

2.5. Evaluation metrics

2.5.1. Accuracy. To test the accuracy of the resulting model, the model is given a new set of mixed healthy and abnormal brain MRI images. The test is to let the model generate the segmentation result and evaluate their precision and recall with confusion matrix. In this test, the study does not consider how close does the segmentation is to the actual brain tumor in the MRI images. Instead, only whether or not the model mark one image as the unhealthy one matters.

$$Precision = \frac{TP}{TP+FP} \tag{3}$$

$$Recall = \frac{TP}{TP+FN} \tag{4}$$

2.5.2. Training performance. The training performance is evaluated based on the time it takes to train the model to its best, reaching the minimum loss in validation dataset, in a given machine.

2.5.3. Segmentation performance. The segmentation performance can be evaluated by overlay the predicted mask on the original image and evaluate if the predicted mask properly cover the brain tumor area.

3. Results and discussion

The experimental results demonstrate successful training of the VGG model using transfer learning with the transferred weight as the base model. However, both ResNet and MobileNet models, regardless of whether they had transferred weights from ImageNet or not, were unable to complete the segmentation tasks (Table 1). Despite these limitations, the data still provides insights into the general trend of training performance. As the number of trainable parameters in a model increases, the number of epochs required for training also increases. Moreover, the model with transferred weights from the base model requires

DOI: 10.54254/2755-2721/8/20230197

fewer epochs to achieve optimal results. This is due, in part, to the fact that in transfer learning, the base model's weight is locked and becomes un-trainable, leading to a smaller number of trainable parameters in the weight-transferred model compared to the non-weight-transferred model.

Table 1. Overall training result.

	VGG		ResNet50		MobileNetV2	
	W/ weight	W/o weight	W/ weight	W/o weight	W/ weight	W/o weight
Is training successful?	Yes	No	No	No	No	No
Epoch	22	32	34	40	38	58

The VGG model with weight transfer achieved exceptional accuracy, as evident from the results in Table 2. The model successfully detected all MRI images with unhealthy brain tissue in the test data set. It is important to note that the model's primary objective was to identify whether an abnormality existed in the MRI image, rather than the specific location of the abnormality. During the test phase, the VGG model with weight transfer achieved a precision score of 1.0 and a recall score of 1.0, indicating perfect performance in detecting abnormal brain tissue.

Table 2. VGG W/ weight confusion matrix.

	Positive	Negative
True	229	411
False	0	0

The results of the other models, as shown in Table 3, are not promising. The common theme among these models is that they always classify the given MRI image as unhealthy, resulting in a high false positive rate and low precision, while achieving a recall rate of 100%. Although these models may identify all the images with unhealthy brain tissue, their inability to distinguish healthy images from unhealthy ones renders them impractical for real-world medical applications. A balance between high precision and recall is crucial for an effective and reliable diagnostic tool, and the VGG model with weight transferred appears to be a promising candidate in this regard.

The VGG model with weight transfer demonstrates a satisfactory performance in terms of segmentation tasks (Figure 6). As shown in Figure 6, the model successfully identified the approximate location in eight out of ten MRI inputs with most of the segmentation having a reasonably close shape and size to the labeled masks. In contrast, the VGG model without weight transfer failed to produce meaningful results, only outputting empty segmentation masks. These results suggest that the pre-trained weights from the ImageNet dataset play a critical role in the success of the VGG model for segmentation tasks in the medical domain. Further experiments with a larger dataset and a more diverse range of medical images would be valuable to validate the effectiveness of the VGG model with weight transfer for this task.

Table 3. MobileNetV2 W/ weight confusion matrix.

	Positive	Negative
True	225	0
False	401	0

DOI: 10.54254/2755-2721/8/20230197

The results of the MobileNetV2 model with weight transfer in the segmentation task were disappointing (Figure 7). The model consistently classified the MRI inputs as abnormal without providing clear identification of the location of the abnormal tissue. The segmentation masks it produced lacked obvious highlights. This conclusion seems to contradict the earlier observation that the failed-to-train models always classified the images as abnormal. However, upon closer inspection, the segmentation results of the failed-to-train models showed highlights of abnormal tissue in the corners of the images (Figure 8), while the VGG model with weight transfer did not exhibit this behavior.

The peculiar behavior of the models could be attributed to the gradient descent algorithm resulting in a local minimum, where the highlighted pixels satisfy the requirements of minimizing the loss function. The underlying cause might be the limited number of images, which are less than 3000, not being sufficient to properly train the large CNN network. Even with data augmentation, the training process is not optimal, considering that the smallest network has over 3,000,000 trainable parameters.

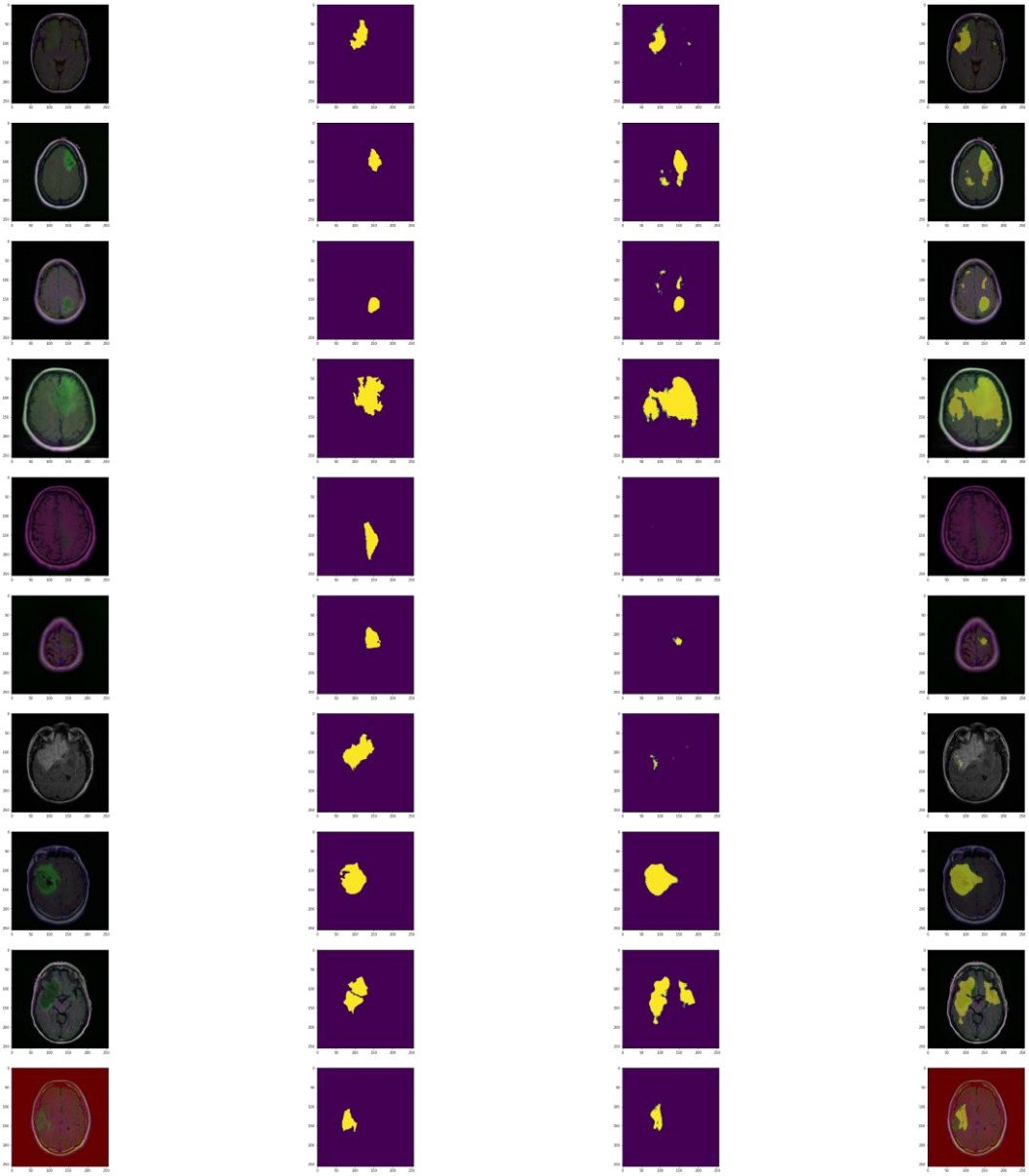


Figure 6. Typical decent segmentation made by VGG W/ weight model. The left most picture is the original MRI image. The middle-left picture is the labeled mask. The middle right picture is the predicted mask. The right most picture is when overlaying preceded image with the orignal image.

DOI: 10.54254/2755-2721/8/20230197

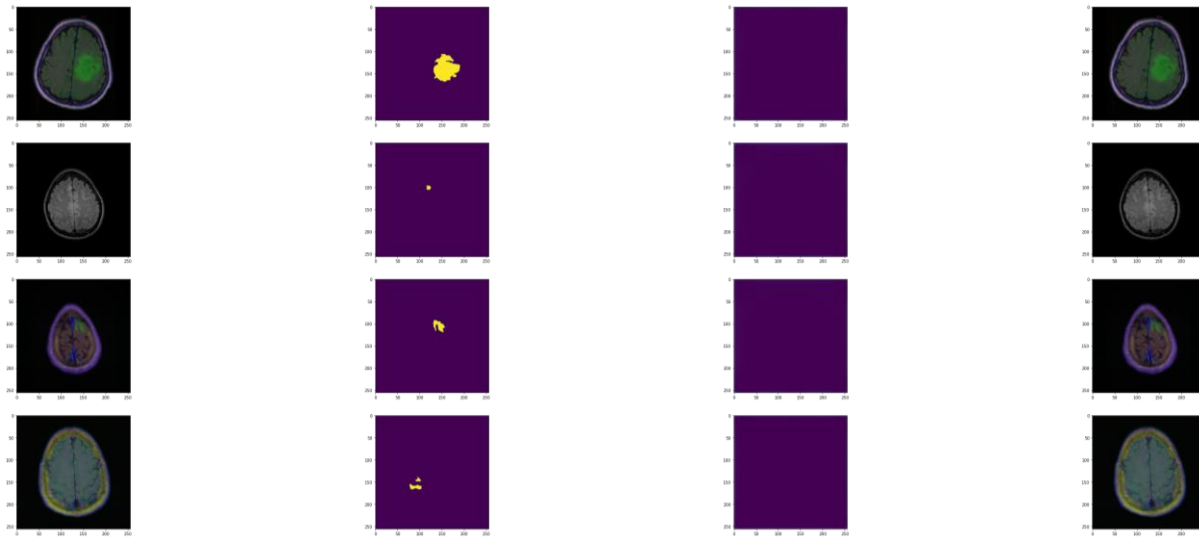


Figure 7. Typical bad segmentation made by MobileNetV2 W/ weight model. The left most picture is the original MRI image. The middle-left picture is the labeled mask. The middle right picture is the predicted mask. The right most picture is when overlaying preceded image with the orignal image.

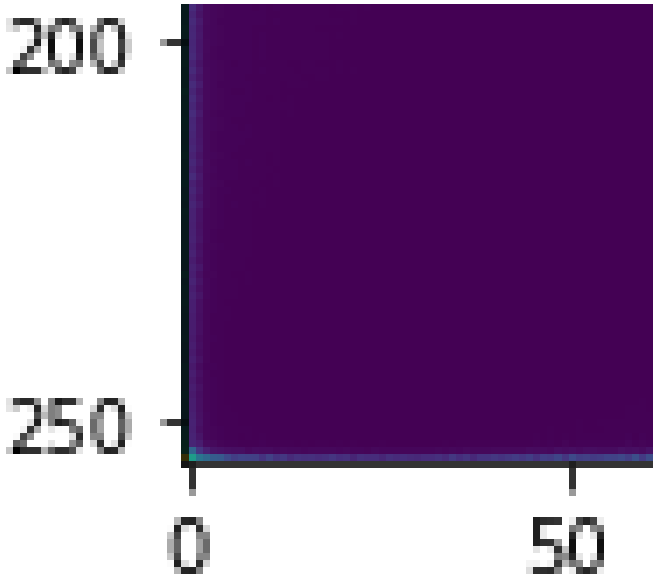


Figure 8. Enlarged image of MobileNet segmentation task.

4. Conclusion

This study aims to verify the transfer learning effectiveness when the original domain is from non-medical domain. It can be concluded that transfer learning with VGG, particularly with weight transfer, is effective in identifying unhealthy brain tissue in MRI images with high accuracy, precision, and recall. On the other hand, transfer learning with MobileNetV2 and ResNet did not perform as well in the segmentation task, with the models always classifying the MRI input as abnormal but failing to identify the exact location of the abnormal tissue. It was also noted that the models with more trainable parameters required more epochs to train and that weight transfer can reduce the number of epochs needed to reach optimal results. However, the limited number of images in the dataset might have affected the optimal training of the larger CNN networks. Overall, the study provides insights into the

DOI: 10.54254/2755-2721/8/20230197

effectiveness of different CNN networks for transfer learning in medical imaging and highlights the importance of dataset size and weight transfer in improving model performance.

References

[1] Cheplygina V de Bruijne M Pluim J P W 2019 Not-so-supervised: a survey of semi-supervised, multi-instance, and transfer learning in medical image analysis Medical image analysis 54: 280-296

[2] Yu H Yang L T Zhang Q et al. 2021 Convolutional neural networks for medical image analysis: state-of-the-art, comparisons, improvement and perspectives Neurocomputing 444: 92-110.

[3] Karimi D Warfield S K Gholipour A 2020 Critical assessment of transfer learning for medical image segmentation with fully convolutional neural networks arXiv preprint arXiv:2006.00356

[4] Wang M Deng W 2018 Deep visual domain adaptation: A survey Neurocomputing 312: 135-153

[5] LeCun Y Boser B Denker J S et al. 1989 Backpropagation applied to handwritten zip code recognition Neural computation 1(4): 541-551

[6] Simonyan K Zisserman A 2014 Very deep convolutional networks for large-scale image recognition arXiv preprint arXiv:1409.1556

[7] Shen D Wu G Suk H I 2017 Deep learning in medical image analysis Annual review of biomedical engineering 19: 221-248

[8] Cheon S Kim J Lim J 2019 The use of deep learning to predict stroke patient mortality International journal of environmental research and public health 16(11): 1876.

[9] Chen S Ma K Zheng Y 2019 Med3d: Transfer learning for 3d medical image analysis arXiv preprint arXiv:1904.00625

[10] Buda M Saha A Mazurowski M A 2019 Association of genomic subtypes of lower-grade gliomas with shape features automatically extracted by a deep learning algorithm Computers in biology and medicine 109: 218-225

[11] Mazurowski M A Clark K Czarnek N M et al. 2017 Radiogenomics of lower-grade glioma: algorithmically-assessed tumor shape is associated with tumor genomic subtypes and patient outcomes in a multi-institutional study with The Cancer Genome Atlas data Journal of neuro-oncology 133: 27-35