2018

Master's thesis

Application of Transfer Learning to 3-Dimensional Medical Imaging

1215087 JIANG, Yunpeng

Advisor Associate Professor. YOSHIDA, Shinichi

March 25, 2019

Informatics Course

Graduate School of Engineering, Kochi University of Technology

Abstract

Application of Transfer Learning to 3-Dimensional Medical Imaging

JIANG, Yunpeng

Nowadays, there are more and more successful deep learning methods which have made remarkable progress in image classification, especially convolutional neural networks(CNNs) have beaten other traditional machine learning methods even human-levels. However, most deep learning model is only used in natural image domain but no other domains, especially only a really small part is used in medical image domain. Magnetic resonance imaging (MRI) is widely used in routine clinical diagnosis and treatments. In Japan, brain dock is being conducted for medical checkup. We get brain dock MRI structured images with gender labels from clinics, and using transfer learning method to apply CNNs on them to estimate the gender. We not only save a lot of doctors' diagnose time, improves clinics' working efficiency, solve the insufficient health care resource problem, but also get a conclusion that there are some relationships between the MRI images with genders. In the future, we can study which part is related to gender and what the relationship is to study human being brain dock structure in neurology and biology. We propose an application of the transfer learning strategy of VGG16 to the brain dock MRI gender estimation and conduct the experiment to clarify the performance. Contributions of this paper are following four works, i.e.,

1. We put forward a new multi-channel fusion method to make the three dimensional grayscale brain dock MRI image available for the pre-trained CNNs. This method

solves the key difficult problem for transfer learning in three dimensional medical image domain.

2. We compare and analyze the difference of strategies between train from scratch and transfer learning in experiment, certify the advantage of transfer learning: avoid the over-fitting problem, reduce a lot of weight parameters and training time, save the computation resource, improve the reuse of models and improve the gender estimation accuracy of brain dock MRI image.

3. We optimize the pre-trained VGG16 model by using fine-tune and fully- convolution methods and using a new method of training learning rate to reduce more weight parameters, get a good model convergence result and improve both the model classification accuracy and speed.

4. We compare and analyze the advantages and disadvantages of different network architecture models, such as: VGG16, VGG19, analyze the successful reason and condition of transfer learning.

The result shows that transfer learning is better than training from scratch and the VGG16 optimized model is best of all. The proposed method and idea of the transfer learning to 3-Dimensional medical imaging can be extended for other medical imaging tasks and also can be useful to make a development in medical diagnosis domain.

key words Transfer learning, Medical imaging, Convolutional neural networks, Computer-aided diagnosis, Image recognition.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

In this chapter, we first introduce our research background and significance. This research is about MRI and application of transfer learning to gender estimation. Then, we overview the research status about Deep learning, Medical imaging and Convolution Neural Networks(CNNs). Finally, we describe the Transfer learning and why we choose it to do our research and experiments.

1.1 Background and Significance

Magnetic resonance imaging (MRI) or computed tomography (CT) is widely used in routine clinical diagnosis and treatment. Nowadays, there are more and more successful deep learning methods which have made remarkable progress in image classification, especially convolutional neural networks(CNNs) have beaten other traditional machine learning methods even human-levels, can help doctors a lot in detection. Deep learning has a great revolution in medical research, for example, detect mitosis in breast cancer cells, predict the toxicity of new drugs, understand gene mutation to prevent disease, computer-aided diagnosis (CADs) and so on.

However, most deep learning model is only used in natural image domain but no other domains, especially only a really small part is used in medical image domain. Magnetic resonance imaging (MRI) is widely used in routine clinical diagnosis and treatments. In Japan, brain dock is being conducted for medical checkup. We get brain dock MRI structured images with gender labels from clinics, and using transfer learning method to apply CNNs on them to estimate the gender. We not only save a lot of doctors' diagnose time, improves clinics' working efficiency, solve the insufficient health care resource problem, but also get a conclusion that there are some relationships between the MRI images with genders. In the future, we can study which part is related to gender and what the relationship is to study human being brain dock structure in neurology and biology. The proposed methods can be extended for other medical imaging tasks. The proposed method and idea of the transfer learning for computer-aided diagnosis can be useful to make a development in medical diagnosis domain.

1.2 Deep Learning

Tracing the history of convolutional neural networks (CNN), the neurocognitive machine proposed by Fukushima in 1980[1] (neocognitron) is considered to be the earliest convolutional neural network model. It is the first multi-layered neural network containing convolution and sub-sampling operations. Between 1989 and 1998, LeCun et al. proposed and improved the LeNet-5 mode [2], which can be used for identifying on handwritten data sets and have high accuracy. It was successfully applied for the commercial use in the field of mail and check for the first time. The convolutional neural network model inherits many of the advantages of the LeNet-5 model. It can be said that LeNet-5 is the originator of modern convolutional neural networks. However, for a long time afterwards, CNN research was stagnant and waited until the birth of deep learning technology to reappear in the researchers' field of vision. In 2006, Hinton et al. used the layer-by-layer training method to use the unsupervised learning method of shallow network to accumulate a deep belief network DBN[3]. As a result, deep learning transitions from the theoretical stage to the practical stage, often appearing in people's field of vision. However, the true climax of deep learning research appeared in the Imagenet large scale visual recognition challenge (ISSVRC) in 2012. Alex Krizhevsky, a student of Hinton, proposed a new deeper deep convolutional neural network model AlexNet[4], which can be considered as a deeper and wider version of LeNet-5. AlexNet first used the new activation function (ReLU) in CNN, the new regularization method, and the method of random ignoring (Dropout) to successfully reduced top-5 (if the real label is contained in the five categories with the highest probability of image classification, it is the correct classification) the error rate to 16.4%, defeating all other contestants to win the championship and achieving an exciting effect. Since then, researchers have begun to pay close attention to and study deep learning, especially convolutional neural networks. The research on image recognition has exploded, and more models and adjustment techniques have emerged, which further enhances CNN's image recognition ability accuracy and training speed. The famous successful models are Overfeat[5], VGGNet[6], Google's GoogLeNet[7, 8, 9], Microsoft's ResNet[10] and so on.

At the same time, in the academic and industrial circles, the rapid development of deep learning has also promoted the numerous achievements of artificial intelligence. In foreign countries, in 2011, Google took the lead in developing the Google brain project integrating deep learning. In 2012, it introduced neural network in applications such as Google Translate. In 2014, it acquired DeepMind and in March 2016 developed AlphaGo based on deep reinforcement learning project[11]. The Go program defeated Li Shishi, the former world champion in the world of Go, and presented artificial intelligence to the eyes of ordinary people for the first time. This is a subversive event in the field of artificial intelligence, which greatly enhances people's interest in artificial intelligence. In the first half of 2011, Microsoft applied deep learning to commercial products such as Bing voice search and X-Box voice commands, and reduced the recognition error rate to 6.3% in September 2016. This has basically reached the human voice recognition ability; in December 2013, FaceBook hired world-renowned artificial intelligence expert Yann LeCun to lead the new AI research laboratory. The neural network translates about 2 billion posts in more than 40 languages every day, which greatly improves the way people communicate. The above three technology giants will develop the entire field of artificial intelligence quickly to an unprecedented height.

1.3 Medical Image Diagnosis

Medical image diagnosis mainly includes X-ray examination, computed tomography (CT), ultrasound, mammography, magnetic resonance imaging (MRI) and positron emission tomography (PET) et al.[12]. The current image diagnosis is mainly diagnosed by a professional radiologist, and preliminary conclusions are drawn, and then submitted to the attending physician for in-depth examination and diagnosis. However, as the number of patients increases and medical conditions develop rapidly, the patient's medical image data also increases exponentially. It is increasingly difficult for professional radiologists to manually meet the needs of patients for timely diagnosis, which poses a huge challenge for subsequent treatment. In order to assist doctors in medical image diagnosis, machine learning and image processing technology have gradually become an emerging development direction and research hotspot in the medical field.

The traditional machine learning image processing technology is mainly divided into three steps, image preprocessing, manual design feature extraction, and classifier classification. The structure is shown in figure 1.1.

The earliest medical image diagnostic technology was X-ray film, which is still one of the most commonly used inspection methods. Nagata et al.[13] used the active contour method in chest X-ray diagnostic nodules and the multi-domain value algorithm was improved. Htike et al.[14] showed a three-layer framework for the diagnosis and

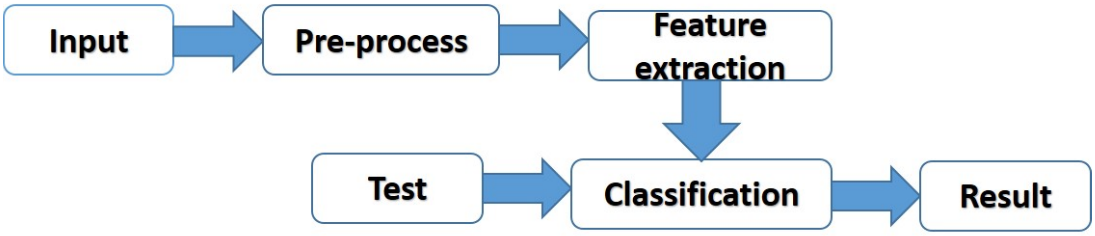


Fig. 1.1 Traditional Machine Learning Image Processing Sample.

treatment of pulmonary nodule images: pretreatment layer, feature extraction layer and classification layer. In the preprocessing layer, the author chose the Laplacian filter. In the feature extraction layer, features such as contrast are extracted from the gray level co-occurrence matrix of the image. In the classification layer, the rotating forest technique was chosen. However, X-ray film is a light and shadow projection technology that can only display the structure of objects in one direction. Because different organs in the human body may overlap more or less, the image display becomes unclear. CT technology overcomes this shortcoming well by projecting multiple angles of the body and synthesizing cross-sectional images to form a series of continuous tomograms. At present, the most effective diagnosis technique for lung cancer is CT diagnosis. The earlier the lesion is found, the smaller the lesion is and the higher cure[15]. Messay et al.[16] used a down sampling algorithm to sample CT images for training, producing comparable inter-slice distances while increasing computational speed and reducing noise. Darmanayagam[17] used the Wiener algorithm to remove additive noise, and Ashwin[18] used the finite contrast adaptive histogram equalization method (CLAHE) to solve the problem of low image contrast. In addition, there are algorithms such as Enhancement Filter, Fast Fourier Transform, Wavelet Transform, Noise Correction[19], Median Filtering, and Gabor Filter. The deep learning model can replace the artificial manual design from the self-learning feature extraction in the image, which improves the diagnostic accuracy of medical images. CNN has a very good performance on the classification task. Yang et al.[20] augmented the LIDC-IDRI dataset[21] and used a 4-layer convolutional neural network model for binary classification identification. The best classification error is less than 0.05. Setio et al.[22] used a 3-layer convolutional neural network with a maximum pooling layer between adjacent layers, which achieved good classification results. However, the deep learning model requires a large enough number of tag data sets for training, so that the picture features can be better extracted to meet the practical diagnosis rate. However, medical image data is generally a limited set of small sample tag data. In order to solve the problem of insufficient data volume and over-fitting, people have begun to try to transfer the CNN network model pre-trained by ImageNet to the field of medical image diagnosis and gotten good experimental results. Xu et al.[23] first adopted the ImageNet pre-trained CNN model, and characterized the glioma pathological images after classification. The images were classified into glioblastoma multiform (GBM) and low grade glioma (LGG). Hoo-Chang Shin et al.[24] transferred the deep learning pre-training model to the diagnosis of lymph nodes and interstitial lung disease in the thoracic and abdomen domain, and obtained a higher recognition accuracy rate, and analyzed the reasons for the success of transfer learning.

In general, in the medical image diagnosis, the successful application of transfer learning to the deep learning model exceeds the traditional image recognition technology, which improves the recognition accuracy. It greatly reduced the doctor's diagnosis time, made full use of the limited resources of the hospital, and effectively alleviated the shortage of medical resources.

1.4 Convolution Neural Networks

1.4.1 Basic Composition

Convolutional neural networks generally consist of four parts. The first part is the convolution layer(Convolution), the second part is the activation function (ReLU), the third part is the pooling layer (Pooling), the fourth part is the full connection layer. First, the input picture is passed through the convolution kernel to make the feature extraction. Then, the feature is extracted by the pooling layer by down sampling to obtain the feature map. Finally, the full connection layer is connected to combine all the local features into global features to calculate the output score of each class, and the input is mapped to the output to obtain the final classification result.

1.4.2 AlexNet

AlexNet is the foundation stone of modern deep CNN. In 2012, Hinton student Alex Krizhevsky proposed the deep CNN model AlexNet, which is widely considered to be a deeper and wider version of LeNet-5. AlexNet first used the new activation function (ReLU) in CNN, the new regularization method to randomly ignore (Dropout), and the use of two GPUs for parallel acceleration operations and other methods successfully reduced the top-5 recognition error rate to 16.4%. It has a huge improvement compared with the runner-up's 26.2% error rate, defeating all other contestants to win the championship and achieving exciting results. On the network architecture, AlexNet has 60 million training weight parameters and 650,000 neurons, the convolutional layer connects to the largest pooling layer, and finally the top layer connects 3 fully connected layers.

AlexNet has the following three advantages: 1. In terms of activation function, although the ReLU activation function has been proposed for a long time, it was not until AlexNet successfully using ReLU as the activation function of CNN, and verified that the effect in the deep network is better than the sigmoid activation function, and the gradient diffusion problems of sigmoid is successfully avoided to able to continue training and learning.

2. In terms of overfitting problems, use Dropout to randomly ignore a part of the neurons to form a small neural network to avoid problems such as over-fitting of the model. Using data enhancement technology, the 224×224 size region is randomly intercepted from the original image of 256×256 and the basic processing of the image is performed by horizontal flipping. The large amount of data is added to effectively solve the CNN over-fitting problem with many parameters. It greatly improves the generalization ability of the model.

3. Use overlapping maximum pooling in CNN. Traditional CNN generally uses average pooling to do down sampling operations, but AlexNet uses maximum pooling to do down sampling operations, which is better than the blurring effect produced by average pooling.

1.4.3 VGGNet

In 2014, the Visual Geometry Group VGG and Google Deep Mind jointly developed VGGNet. By analyzing the depth and performance of the convolutional neural network, the size of the convolution kernel was reduced and multiple 3×3 were used repeatedly. The convolutional neural network architecture with a depth of 16-19 layers was successfully constructed by the operation of the convolution kernel of size and the largest pooling layer of 2×2 size. Compared with the previous state-of-the-art CNN network architecture model, VGGNet has significantly reduced the recognition error rate and achieved the amazing result of the runner-up and positioning project champion of the ILSVRC 2014 Challenge image classification project. The network architecture is

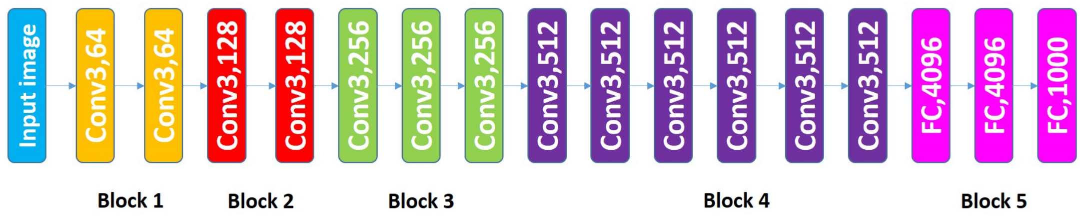


Fig. 1.2 VGG16 Networks Architecture.

shown in figure 1.2.

1.4.4 Disadvantages

Both the traditional CNN convolutional neural network architecture and the classic network architecture described above face a key core issue, namely a large number of tagged training data and a sufficient set of validation and test sets. When the amount of data is insufficient, there will be a serious problem of overfitting, which results in the recognition of the model on the training set is very good, and sometimes even close to 100%, but the accuracy of the test set is much lower, indicating The model is difficult to generalize to the general situation, such a model cannot complete the picture recognition task. In daily life, it is difficult to obtain sufficient tag data, especially privacy data such as medical images. Therefore, we began to study transfer learning, trying to retrain data in different fields using pre-trained models, which not only reduces the complexity of developing new models, but also solves problems such as over-fitting. It can also make models be able to reused multiple times to increases utilization of effective resources instead of reconceiving design for each new problem and saving existing resources.

1.5 Transfer Learning

1.5.1 Definition and Classification

Traditional machine learning[25] usually has two basic assumptions. The training sample and the test sample satisfy the independent and identical distribution, and there must be enough training samples available. However, real life is that these two basic assumptions are difficult to satisfy at the same time. Transfer learning, which has received widespread attention in the field of machine learning in recent years, offers the possibility of solving these two problems. Transfer learning solves the problem of insufficient data volume of small sample tags in the target domain by using existing knowledge, thus fundamentally relaxing the basic assumptions of traditional machine learning. Transfer learning also has the unique wisdom of human beings. It can transfer models suitable for large-label large data sets to small data sets. This solves the problem of over-fitting and enables deep learning in different fields. More than just natural image recognition, it broadens the scope of application and improves the use of effective resources[26].

Transfer learning consists of two parts, a domain and a task. The domain is divided into a source domain and a target domain, the tasks are divided into source tasks and target tasks. Target and source domains, target and source tasks are sometimes the same and sometimes different. Taking the MRI medical image diagnosis of Brain Dock as an example, the source domain is ImageNet natural image data, the target domain is Brain Dock MRI image data, the source domain and the target domain are different, the source task is to classify the image into 1000 classes, but the target task is to classify the image into 2 classes of gender, so the source task and the target task are different. So in the field of deep learning, we generally use parameter knowledge transfer to share model parameters. The original pre-trained model weight parameters are transferred to the

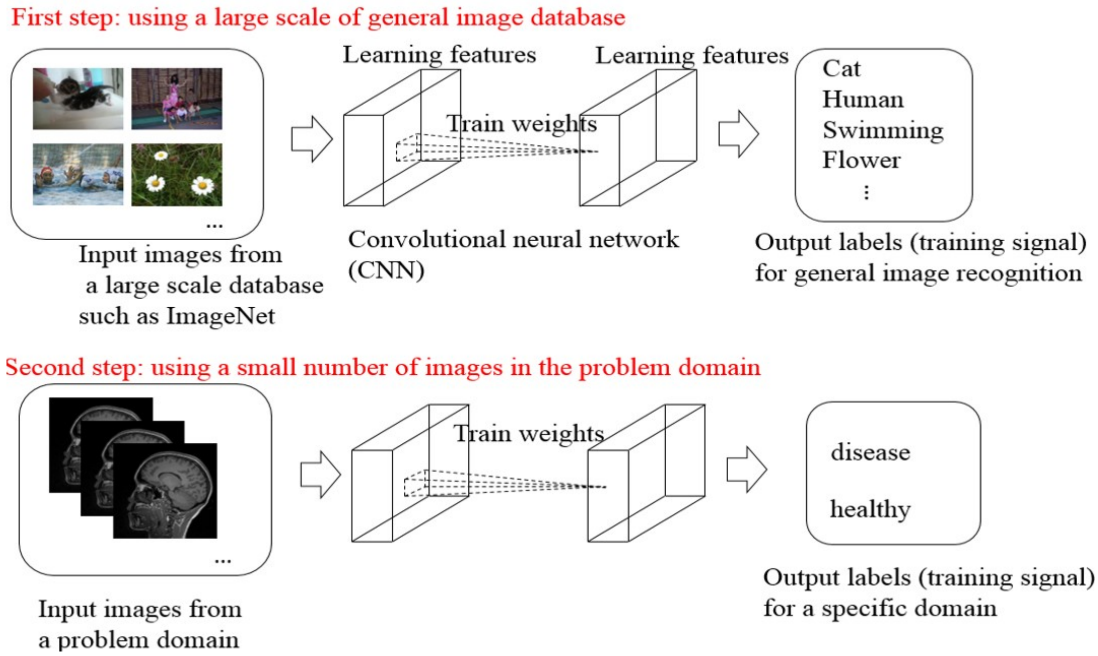


Fig. 1.3 Transfer learning Steps.

target domain. The figure 1.3 shows that: According to the research of human visual neural network, the features extracted in the low-level hierarchy are basically the same, such as shape, color, edge, etc., but the features extracted in the high-level hierarchy are very different. So we can directly extract the underlying model convolution kernel weight parameters for feature extraction in the data transfer. Because of the large number of training sets and excellent verification set recognition accuracy, such convolution kernels use a smaller number of targets instead, and the convolution kernel from the picture training is even better. The features extracted from the model convolution kernel at the top layer deepen with the depth of the layer, and the extracted features tend to be more and more oriented to the target domain image features. So we need to retrain the target domain image to get the target image features such as brain dock MRI features in order to achieve higher recognition accuracy.

1.5.2 Advantages

At the widely acclaimed NIPS 2016, former Baidu chief scientist and Stanford University professor Andrew Ng said that transfer learning will be the driving force behind the next machine learning business success after supervised learning, which pushed itself to a climax. Medical image diagnosis cannot directly use deep learning because it cannot obtain a large number of tagged data training sets, and it is prone to over-fitting and other problems. The emergence of transfer learning can solve the over-fitting problem of small sample data sets. Therefore, it can assist doctors in medical image diagnosis, improve doctors' work efficiency, and make full use of limited medical resources.

Traditional image recognition requires manual design feature extraction. This work requires a large number of domain knowledge experts to perform manual design and experiments according to specific image features to extract corresponding features [38]. This process requires a lot of experimentation and time to verify, but the extracted features often do not meet the recognition requirements or even some simple feature extraction. However, deep learning through end-to-end image recognition, self-learning feature extraction instead of manual design feature extraction, not only reduces a large amount of design time and labor costs, reduces the requirements of domain expertise, and the characteristics of learning equal to or better than the characteristics of the artificial design. We only need to input the image into the convolutional neural network, and we can directly obtain the classification result to diagnose the disease, greatly reduce the difficulty and complexity of the recognition task, and improve the accuracy and diagnostic efficiency of the recognition. This enables early diagnosis and treatment, control of disease progression, and even save lives.

Through transfer learning, the excellent network model that has been pre-trained is applied in the medical field. The bottom layer can also obtain the basic features such as the edge and texture of the image, and the top layer can obtain better features by training the data in the medical field. This not only solves the problem that the amount of medical data is small and easy to over-fitting, but also improves the training speed and accuracy, and provides a convenient method for solving similar problems, which broadens the application range of the model and improves the effective utilization rate.

Chapter 2

Brain Dock MRI Imaging Processing

In this chapter, we describe our experiment dataset and related works. First, we introduce our brain dock MRI image information. Then, we use python to read and perform several pre-process such as normalization and put the data into the list to composite the datasets. We use data augmentation to solve the class bias problem and get more data to avoid over-fitting problem to make our model have a high recognition accuracy and a good balance in the next training models steps. Finally, we propose a new method called Multi-channel fusion to make the grayscale medical imaging colorful and available to the VGGNet model.

2.1 Dataset Introduction

In Japan, brain dock is being conducted. Brain dock is a kind of medical checkup focused on MRI data. We get some sample images which already eliminate all the private information from a clinic with permission. It is MRI structured image, Imaging the structure of each brain tissue with grayscale 3 dimensional data which has so many slices. It has some labels such as, gender, age, cigarette, alcohol, and so on. We can use this to train the model to predict people ’ s future health. In the first step, gender is so important in brain dock project because the transgender is a serious problem in

Fig. 2.1 Brain Dock MRI Sample.

the future. The figure 2.1 shows that: There are 364 labeled MRI images including 157 females (Negative) and 207 males(Positive), which will cause a serious class bias problem later. Each file represents one patient, contains a series with multiple axial slices of brain dock MRI. Each slice image size is 256×256, and format is NIFTI.

2.2 Data Pre-process

First, we use Python's nibabel library to read the NIFTI format image, call the operating system command to sequentially read the folders in the dataset from the folder root directory, each folder is a MRI scan image of the patient. We read all the MRI images with all slices, make the normalization and put them into the list in order. Then, we use python's numpy library call function to read the .csv file, because the first line

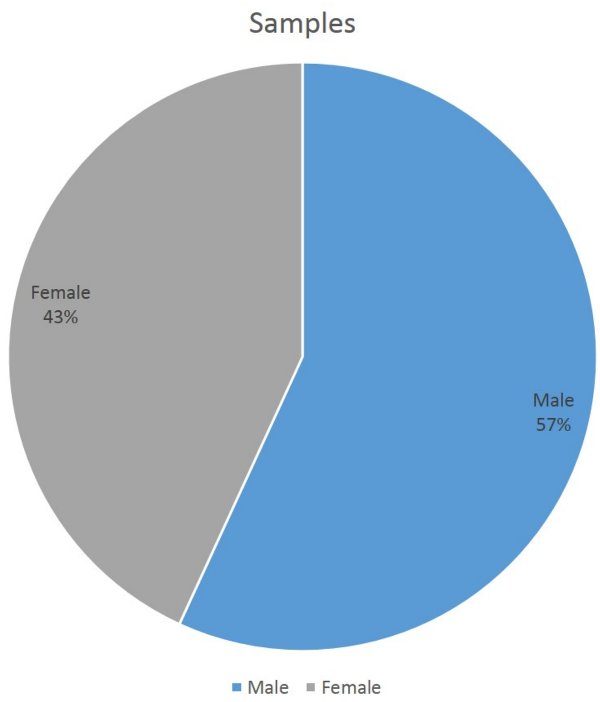


Fig. 2.2 Class Bias Sample.

is useless information, the first column is id value does not work, so we read directly from the second line, and because we only need to get the label, so read the second column directly into the list. Through the above operations, we obtained preliminary MRI images and labels to prepare for subsequent operations.

2.3 Data Augmentation

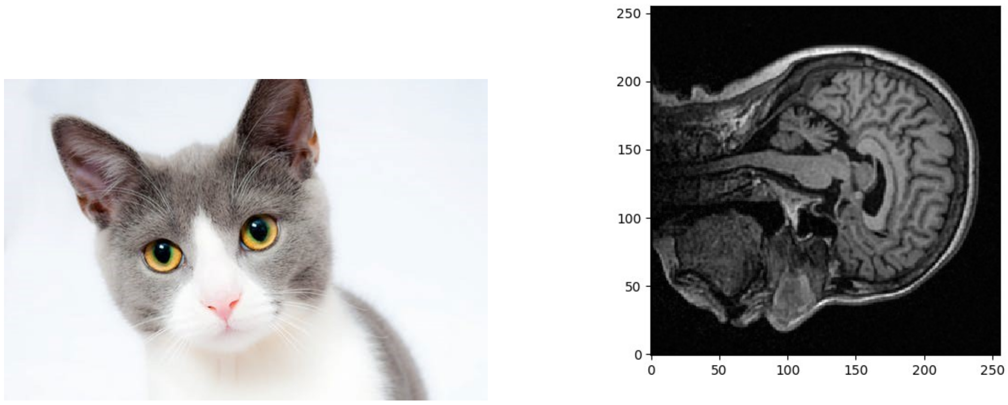
In order to better train the effect, generally our data set requires a balance of category distribution which the number of samples of each class is approximately same. If the number of samples in a certain category is too large or too small, this will result in a prediction probability that is far superior to or far lower than that of other categories, called class bias problem. Especially in the field of medical image diagnosis, patients with clinical manifestations and other test data, after the initial diagnosis results have been obtained in order to obtain more powerful evidence to prove the rationality of the diagnosis before the MRI scan diagnosis. In this way, the data distribution obtained is always seriously unbalanced. For example, in our experiment, our brain dock MRI images with gender label which male is much more than female as the figure 2.2 shows. If we do not deal with the data set, it will result in a prediction of gender, the probability of positive samples (Male)is much higher than the probability of negative samples(Female). This will seriously affect the recognition accuracy of our model. We have adopted data augmentation techniques to solve the problem of class bias.

We use the data augmentation method to oversample the positive samples by random 3 times and oversample the negative samples by random 4 times to make class balance and more data samples. Now, there are 1249 samples including 628 negative (Female) samples and 621 positive (Male) samples so that we have a good balance and more samples data collection to train the model in the next step.

2.4 Multi-Channel Fusion

There is a big difference between natural image and medical image especially MRI image in NIFTI format which is used to train CNN model as input. Natural image is 3D colorful image which has RGB channels, but NIFTI image is a series of 2D grayscale image slices. figure 2.3(a) shows that the natural colorful image sample, and figure 2.3(b) shows that the brain dock MRI NIFTI grayscale image slice sample. There is a big difference between them as you can see in the figure.

VGG16 model is used for natural image so the model needs 3D colorful image input. We put forward a new multi-channel fusion method to make it available for VGG16 input.



(a) Natural Colorful Image Sample. (b) Medical Imaging MRI Sample.

Fig. 2.3 Natural Image and Medical Image Differences.

2.5 Fake 2.5-Dimension

We choose all 3 slices of the image from numpy list, concatenate them with each other to represent multiple fake 2.5-Dimension images with 3 channels. Because the natural image is 3 channels with RGB so that we can see the red, green, blue different colors to see the real colorful image by our eyes. However, our image only has 3 channels which are actually different slices of grayscale MRI image but not the real RGB channels so that we call it fake 2.5-Dimension images which can be only available for CNNs input. The figure 2.4 shows that.

We do a lot of experiments and read papers to find which 3 slices are more suitable

Fig. 2.4 Multi-Channel Fusion Steps.

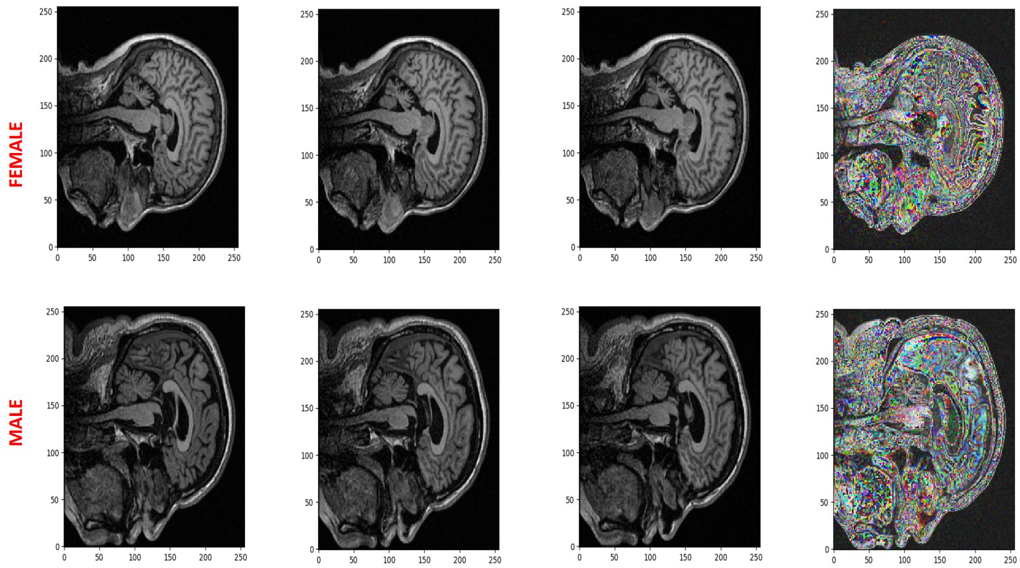


Fig. 2.5 Fake 2.5-Dimension Image Results.

and find that the middles are best because they are the axial and the result is best of all. The first 3 slices represent "RGB" channels and we concatenate them with each other to get a fake 2.5-Dimension image as you can see in the last "colorful" picture. The figure 2.5 shows that.

Chapter 3

Three-Dimensional Medical Imaging Diagnosis Methods

In this chapter, we describe 3 methods of our gender estimation experiment. First, we compare the 2 methods: Train from Scratch and Transfer Learning and certify that Transfer Learning is better in both accuracy and speed because it avoids over-fitting problem of common insufficient labeled dataset. Finally, we use this model as basic model and optimize it by using fine-tune, full convolution methods to improve our model architecture and gradually change the learning rate to make the model get fast convergence and get higher recognition accuracy.

3.1 Train from Scratch

In our experiment, we choose VGG16 to train from scratch directly with brain dock MRI labeled data. We read a lot of papers find that medical image is not suitable for either much lower layers like AlexNet, because it will cause under fitting problem or

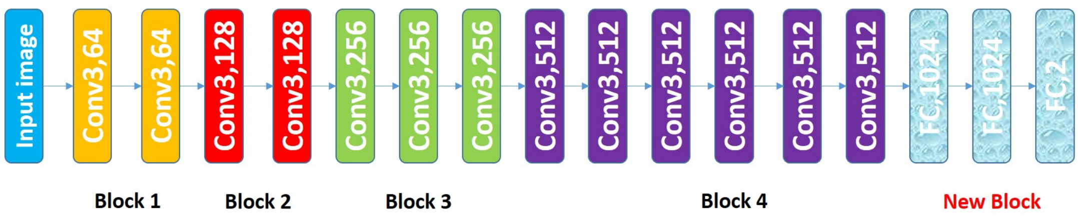


Fig. 3.1 Train from Scratch Architecture.

much deeper layers like GoogLeNet, because it will cause over fitting problem. Because of the different size of the image input, we are unable to use the pre-trained full VGG16 network model architecture. Therefore, we only retain all the convolutional layers and remove the top-level full connection layer. We add a new three-layer full connection layer at the top layer to form a new network model architecture. Because we don ' t have so much big data like ImageNet and we just need to do binary classification of gender, we choose a little smaller layer with just a few weight parameters and the final layer is just with 2 neurons to avoid over fitting problem. The network architecture is shown in figure 3.1.

We used small batch processing to train data since the image size is $256 \times 256 \times 3$, when the batch size is set larger than 16, the memory will be insufficient when it is read into the memory, and it cannot be trained. When the learning rate used is too large, the loss result is difficult to converge. Because in the end we have to perform the two classification, so the binary cross-entropy loss function is adopted. In terms of the optimizer, we used Adam with faster training and better results and adopted default parameters. The initialization parameters are shown in table 3.1. This experiment uses this architecture as a prototype to compare the two training methods: training from scratch and parameters knowledge transfer learning.

Table 3.1 Training Detail Parameters

Batch_size	16
Input_size	256
Epochs	380
Loss	Binary_crossentropy
Optimizer	Adam(beta_1=0.9,beta_2=0.999,epsilon=1e-08)

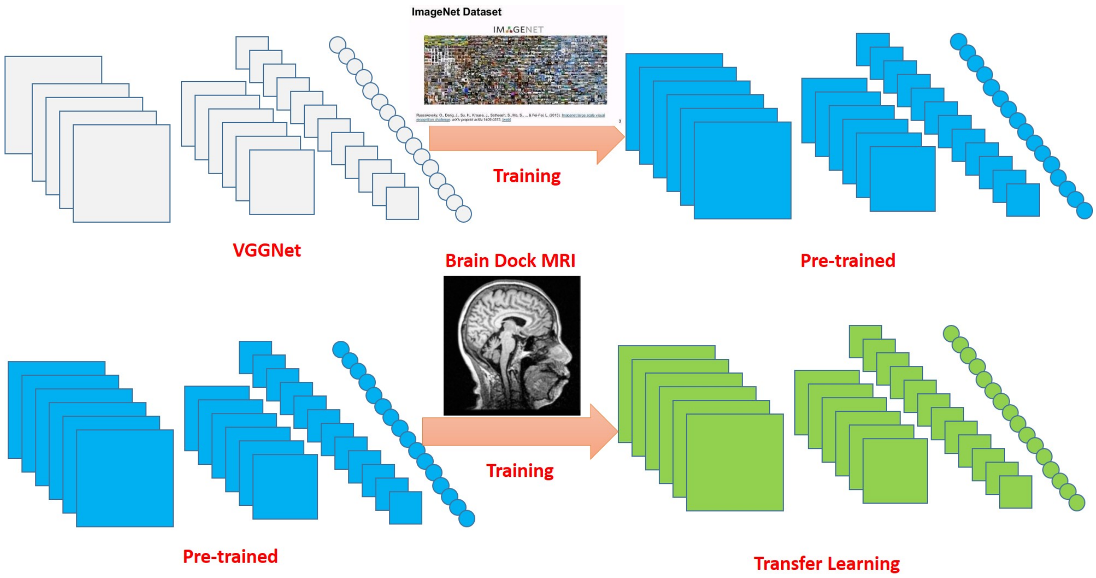


Fig. 3.2 Transfer Learning Architecture.

3.2 Transfer Learning

Generally, transfer learning has 2 steps. First, we use a large image database like ImageNet to train the network to learn the features, output labels and update weight parameters. Then, we transfer the pre-trained weight parameters and train the model with our medical domain images to learn the features in the low layer and output our labels in the top layers to avoid over fitting problem. It not only reuses the pre-trained model and make the training speed faster, but also converge earlier and get higher recognition accuracy. For example, the transfer learning for brain dock MRI diagnosis process is showed in figure 3.2.

We use parameters knowledge transfer learning method to share the weight parameters. We use the VGG16 pre-trained weight parameters because lots of experiments show that we can get the same features in low layers such as: color, edge, but get different features in high layers. The network architecture and the initialization parameters are the same as training from scratch.

3.3 Optimize Model

3.3.1 Fine-tune

We fine-tune the transfer learning model by freezing the low layers but train the top layers to update weight parameters. In order to find the number of best frozen layers, we do a lot of experiments from layer 1 to layer 13 each layer 4 times and take the average results. When we freeze the low layers, for example,1-4 layers, we find the model is over-fitting that it can recognize the training set very well but very poor in validation set so that the validation loss is increasing because we don ' t have much enough amount data to train the model. With the number of layers increasing, the model avoids over-fitting problem and works well in both training set and validation set because we freeze some low layers to reduce weight parameters and we can get good common image features like: edges by using shared weight parameters from ImageNet. But when the number of layers is larger than 7, the model loss is increasing rapidly because we freeze so many layers to get the high features which are much different from the brain dock MRI image features. There are so many differences between natural images and medical images, the pre-trained VGG16 model weight parameters are trained to get the high features of natural image so we can ' t use the high layer weight parameters to get brain dock MRI image features directly. We find the best frozen layer is 7 layers both in experiment results and logical analogies as mentioned. The result is consistent with the expected idea: The lower layer can get the same features, but the higher layers can get different features so the middle layer is the best choice [7].

3.3.2 Full Convolution

The tradition CNN model has a flatten operation to get the feature map and connect to the full connection layer to get the classification result. There are two disadvantages

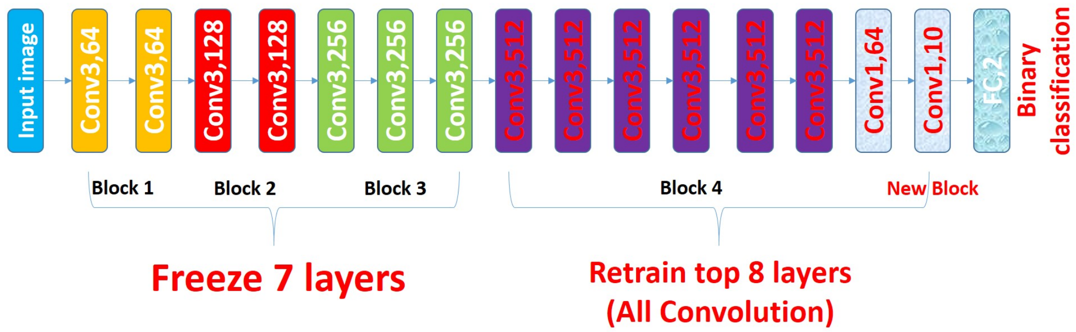


Fig. 3.3 Optimized Model Architecture.

in the model: 1. There are more weight parameters which increase training time and cost much of the computation resource so that some hardware is not available. 2. There is location information in image, but the flatten operation only make the feature into vector so that lost a lot of location information which is crucial to the classification result [8]. We use fully 1×1 convolution to replace full connection not only to reduce the dimension of features to reduce a lot of weight parameters to reduce computation but also get higher fusion features with location information to get better classification result. The final optimized model architecture is like figure 3.3.

3.3.3 Gradually change the learning rate

The traditional training method is that the learning rate is always the same. Recently, it has been suggested that the learning rate should be gradually changed, but this method has not been used in migration learning. Using the method of gradually changing the learning rate not only accelerates the network convergence, but also solves the problem that the loss value is oscillating and difficult to converge, and the learning weight is gradually reduced. In the top-level network training, better weight parameters are learned. To this end, we conducted a number of experiments to accelerate the convergence of the model and improve the recognition accuracy of the network.

Chapter 4

Brain Dock MRI Gender Estimation

In this chapter, we introduce our experiment details and results. First, we have a brief introduction about gender estimation and experiment hardware and software situation. Then, we compare the 2 different models of VGGNet: VGG16 and VGG19 to analyze the condition of transfer learning and certify that VGG16 is better than VGG19 both in speed and recognition accuracy because it doesn't need so many layers, so many parameters which will cause a over-fitting problem and increase the training time. Finally, we give our experiment results in F1-Score and Confusion Matrix and the visualization of feature maps through different layers to give a direct view and opinion of result.

4.1 Gender Estimation

We want to use brain dock MRI labeled images to make the gender estimation and also want to know if there are any differences between Male and Female brain dock MRI. We not only save a lot of doctors' diagnose time, improves clinics' working efficiency, solve the insufficient health care resource problem, but also get a conclusion that there are some relationships between the MRI images with genders. In the future, we can study which part is related to gender and what the relationship is to study

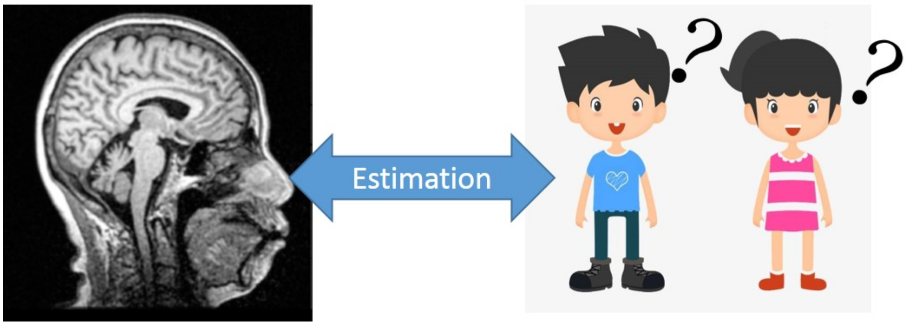


Fig. 4.1 Gender Estimation.

human being brain dock structure in neurology and biology as the figure 4.1 shows. We use 1249 sample images dataset including 628 females and 621 males after processing the original data. In order to train the model well, we spilt the dataset into 2 parts train dataset and validation dataset as 8:2. Finally, we get 999 sample images to train, and 250 sample images to validate.

Table 4.1 Experiment Situation

Process :	Intel(R)Core(TM) i7
Graphics :	NVIDIA GTX 1080Ti 11G
Memory :	32G
Operating System :	Ubuntu 16.04
Programming Language :	Python 3.6
Software Library :	Tensorflow, Keras

4.2 Experiment Situation

Keras is a high-level neural network API based on Tensorflow, written in Python, which shortens the time to validate experimental results and quickly converts ideas into results, greatly simplifying our programming and experimental design process. Keras is able to build simple and fast models, support network architectures such as CNN and RNN, seamlessly switch CPUs and GPUs, and the interface is very user-friendly, enabling researchers to get started quickly. It has the advantages of modularity and scalability.The experimental environment is shown in table 4.1.

4.3 Method Comparison

4.3.1 Train from scratch Vs Transfer Learning

We evaluate and compare the performances of two methods: training from scratch and transfer learning to certify that transfer learning is better and useful. Transfer learning method is better than training from scratch method both in accuracy and speed. There are several reasons: 1. There is limited amount of samples to train the model so that it is over-fitting.

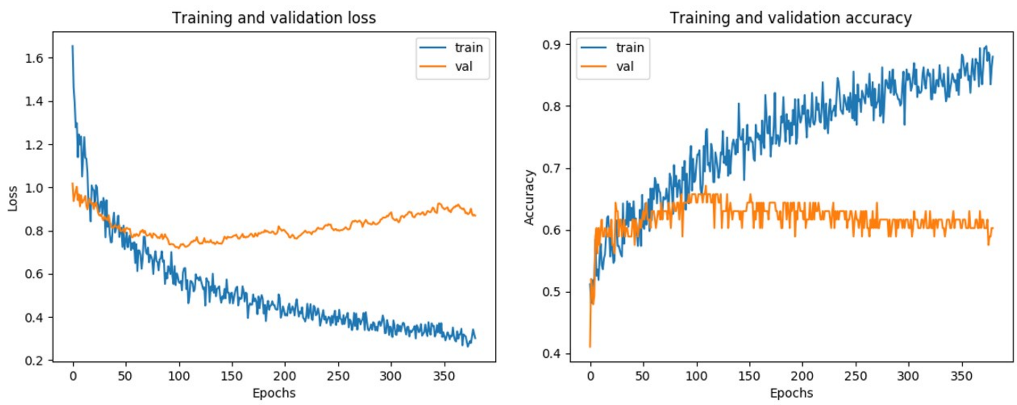


Fig. 4.2 Train from Scratch Results.

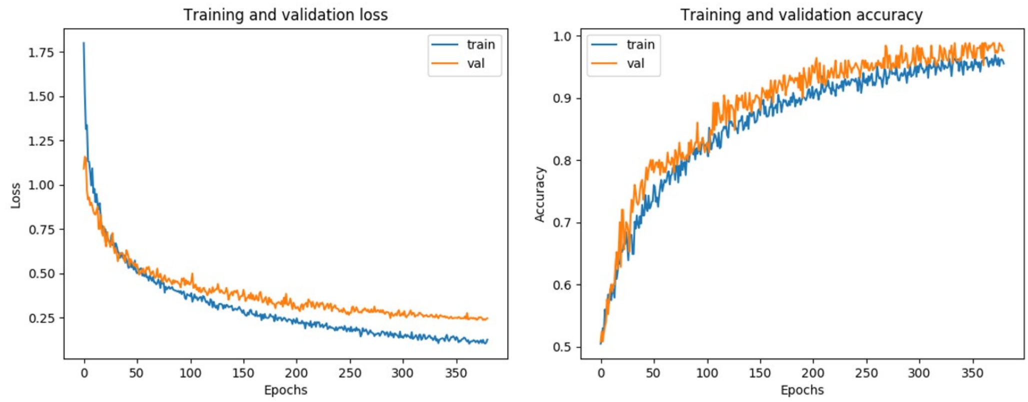


Fig. 4.3 Transfer Learning Results.

The figure 4.2. shows that the training loss is decreasing with the increasing epochs but validation loss is decreasing first and increasing later which shows that the model is over-fitting to the train set so that it gets worse in validation set. The figure 4.3. shows that the training loss is decreasing with the increasing epochs together with validation loss so that it can certifies that the shared weight parameters transfer learning can avoid over fitting problem and it works well both on training set and validation set. 2. The less weight parameters, the less computation needs so that we can speed our training time.

4.3.2 Transfer Learning Vs Optimized Model

We evaluate and compare the performances of two methods: Transfer Learning and Optimized Model and certify that optimized model is better and useful. Optimized model is better than transfer learning method both in accuracy and speed. The image shows that optimized model training loss and validation loss decrease faster and smoother, and training accuracy and validation accuracy increase faster and higher. The figure 4.4 shows that.

4.4 VGG16 Vs VGG19

We have already certified that our final optimized model is best in loss, accuracy, and speed. Next, we will do the experiments about different network architectures. Because we need to do the comparison, we just choose the different VGGNet such as VGG19. Comparison of network architecture: VGG16 is better than VGG19 because the network of 19 is too deep, which leads to the difference between the characteristics of the migration and the target MRI characteristics, resulting in a decrease in accuracy. And because of the small amount of data, there is an overfitting problem. As with the previous IEEE MEDICAL IMAGE paper, medical images should use a network with fewer layers, and the effect will be better. Using too deep a network will reduce the accuracy, the results are shown in figure 4.5.

We evaluate and compare the two optimized models: VGG16 Optimized Model and VGG19 Optimized Model and certify that VGG19 Optimized Model is better and useful. VGG16 Optimized Model is better than VGG19 Optimized Model both in accuracy and speed. The image shows that VGG16 Optimized Model training loss and validation loss decrease faster and smoother, and training accuracy and validation accuracy increase faster and higher.

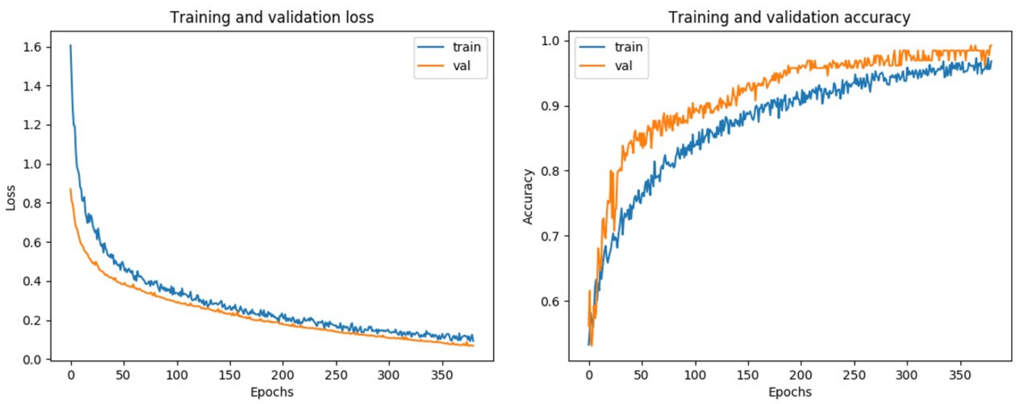


Fig. 4.4 VGG16 Optimized Model Results.

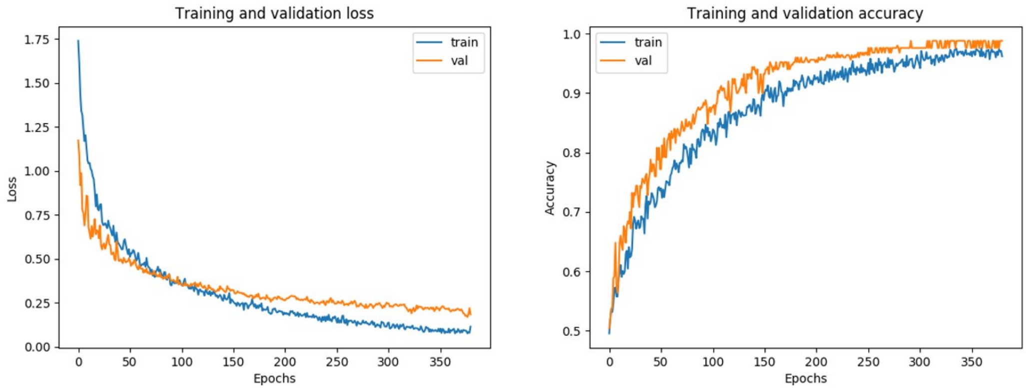


Fig. 4.5 VGG19 Optimized Model Results.

Table 4.2 Compare Methods : TrainSet Loss, Accuracy, Training Time

Model	Train loss	Train acc	Time/epoch
Train from scratch	0.3020	0.8797	325ms
Transfer Learning	0.1260	0.9550	268ms
VGG16 (Fine-tune)	0.1153	0.9644	207ms
VGG16 (Fully Conv)	0.0933	0.9682	186ms
VGG19(Optimized)	0.0764	0.9770	190ms

In conclusion, we do all the experiments to compare all 3 methods and 2 different network architectures of VGGNet, the TrainSet result is in table 4.2 and ValidationSet result is in table 4.3. We can see that our final optimized model is best of all.

4.5 Results

4.5.1 F1-Score

The F1-Score is a measure of the classification problem. In some machine learning competitions with multiple classification problems, F1-score is often used as the final evaluation method. It is the harmonic mean of the accuracy rate and the recall rate,

Table 4.3 Compare Methods : Validation Loss, Accuracy

Model	Validation loss	Validation acc
Train from scratch	0.8695	0.6027
Transfer Learning	0.2465	0.9760
VGG16 (Fine-tune)	0.0720	0.9769
VGG16 (Fully Conv)	0.0686	0.9923
VGG19(Optimized)	0.1267	0.9760

with a maximum of 1 and a minimum of 0.Precision: refers to the proportion of positive samples in the positive example determined by the classifier, Recall: refers to the proportion of the total positive case that is predicted to be positive.True Positive (TP) means making a positive decision and the decision is correct. Therefore, the value of TP indicates the number of correct Positive determinations. Similarly, the False Positive (TP) value indicates the number of incorrect Positive decisions. Accordingly, the True Negative (TN) value indicates the number of correct Negative decisions. The False Negative (FN) value indicates the number of negative decisions. The formulation is listed as below.

$$\mathrm{F}_1 = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \tag{4.1}$$

$$\text{precision} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{4.2}$$

$$\text{recall} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{4.3}$$

The experimental results show that the optimal model is not only close to 100% in accuracy. In PR verification, the effect is also very good. In the F1-Score test, 97.39% of the high score is also obtained, indicating that the model has better processing results for the class bias problem. The model is very stable and result is shown in figure 4.6.

The model predicts class balance, really works well and get a high F1-Score, the

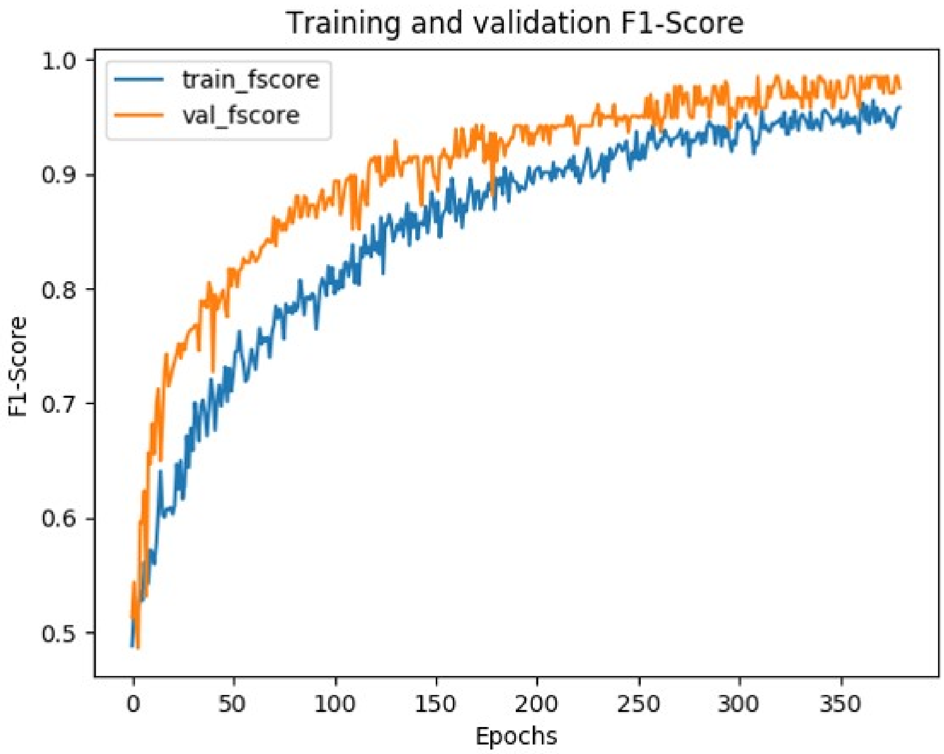


Fig. 4.6 F1-Score Result.

Table 4.4 F1-Score Result.

Final Optimized	Precision	Recall	F1-Score
Train	0.9687	0.9647	0.9648
Validation	0.9792	0.9726	0.9739

detail is shown in table 4.4.

4.5.2 Confusion Matrix

In machine learning, especially in statistical classification, the confusion matrix is also called the error matrix. Each list of matrices reaches the classifier's category prediction for the sample, and each row of the second matrix expresses the real category to which the version belongs. It is a two-row, two-column table that reports the number of four prediction-related events: False Positive, False Negative, True positive, and True negative. This table allows us to analyze the performance of the prediction system in more detail. The heat is not just an accuracy rate. Accuracy is an unreliable classifier performance metric because it can produce misleading results when the number of samples in different classes in the data set is unevenly distributed. For example, if there are 95 males and 5 females in the data set, the classifier will simply divide it into males, which is 95% accurate.

We have 250 sample images in cross validation and the result is shown in figure 4.7. As you can see in the picture, there are over 100 samples are TN which means true label is 0(Negative), and prediction label is 0(Negative) too, it shows that he is a female, and over 120 samples are TP which means true label is 1(Positive), and prediction label is 1 (Positive)too, it shows that he is a male and only few samples prediction is wrong. Our model has a good classification to solve the class bias problem and also has a high recognition accuracy.

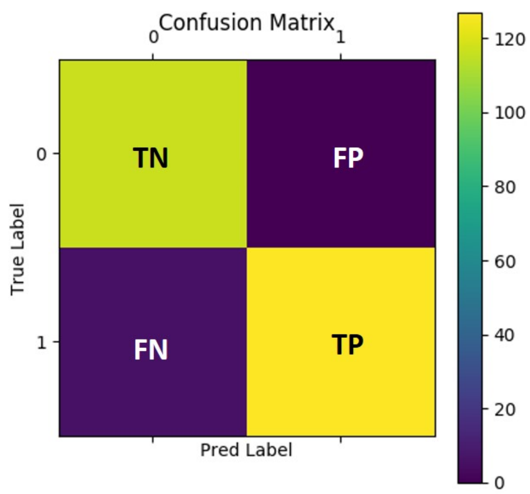


Fig. 4.7 Confusion Matrix Result. label 0 : Female, 1 : Male.

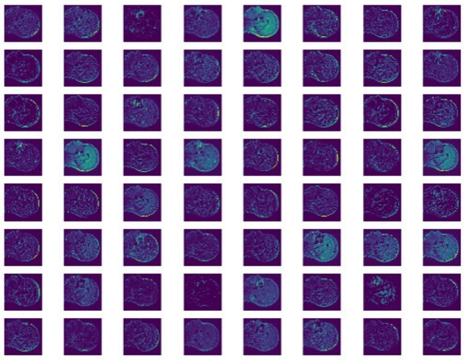


Fig. 4.8 First Layer Feature Map.

4.5.3 Feature Map

We put the brain dock MRI images into the different layers of model, and visualize the feature maps to know more about the image so we just pick 64 kernels and 5 convolution layers to make a visualization. Feature map of the first layer after convolution visualization is shown in figure 4.8.

You can see 64 feature maps by different kernels to make the feature extraction. All the first feature maps are 1:1 integrated after the feature map is shown in figure 4.9. It is similar to our original MRI image, but there are much differences in the following

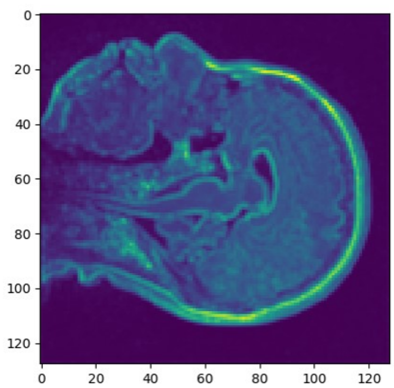


Fig. 4.9 First Layer Integrated Feature Map.

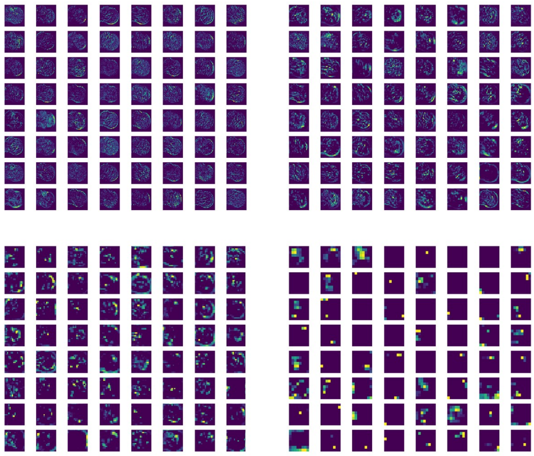


Fig. 4.10 Other Layer Feature Maps.

feature maps.

Feature map of the other 2-5 layers after convolution visualization is shown in figure 4.10:

It gets much different smaller feature maps such as edges, colors, and small part of different compositions. In the final layer, you can see only just small feature extraction which we used to make classification. All the other 2-5 feature maps are 1:1 integrated

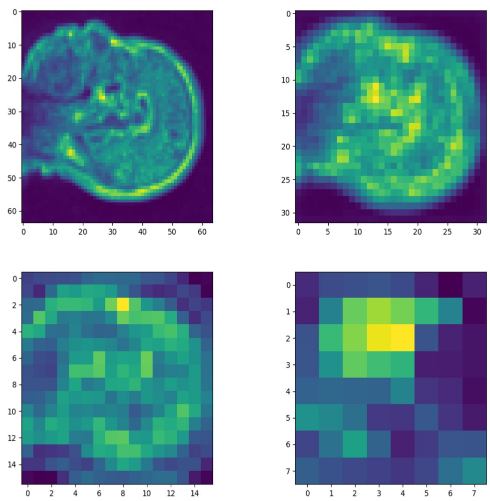


Fig. 4.11 Other Layer Integrated Feature Maps.

after the feature map is shown in figure 4.11:

The first 2 image looks a little clear, but the last 2 image looks only a small part of pixels which means the feature extraction. As you can see, maybe there is something relationship between gender with the core part in brain dock MRI, so that we can make the model have a good prediction of gender and has a high recognition accuracy and also high F1-Score to solve the class bias problem.

Maybe the middle part of brain is something related to gender estimation like the yellow pixels' part in the image, we will study this relationship and try to find whether there is any theoretical foundation in biology and neurology in the future.

Chapter 5

Conclusions and Prospects

In this chapter, we focus on our research conclusions and prospects. We give a brief conclusion about the research and experiment, our background and significance, our methods and experiment details. And our results and findings. We also find some problems and shortcomings in our research and want to improve them in the future study. We have some prospects of deep learning, transfer learning for medical image diagnosis and we hope we can study them in the future and maybe it can give some advice or inspirations to make a development of transfer learning in other domains.

5.1 Conclusions

First, we have a brief introduction about our experiment background and significance. We also introduce recent research status with deep learning and medical imaging diagnosis and also transfer learning. Then, we talk about Convolution Neural Network basic composition and famous model and also the advantages and disadvantages. Finally, we focus on the dataset and related work, we introduce our dataset distribution and do the process and data augmentation to prepare for next step. We put forward a new Multi-channel fusion method: choose and combine the MRI middle 3 slices into 3 channels to represent a 2.5d fake colorful image with RGB as input like natural images to make VGG16 available. We compare two methods: Train from scratch and Transfer learning, and do experiments to certify that transfer learning is better than learning from scratch both in speed and accuracy. We optimize the pre-trained VGG16 model by using fine-tune and fully- convolution methods to improve the experiment results both in speed and accuracy. We compare 2 different model architecture models: VGG16 and VGG19 to analyze and certify that transfer learning method doesn ' t need so many layers and parameters which may cause over-fitting problems, analyze the successful reason and condition of transfer learning.

The result shows that transfer learning is better than training from scratch and the proposed method and idea of the transfer learning can be extended for other medical imaging tasks of computer-aided diagnosis and also can be useful to make a development in medical diagnosis domain. However, there are still many shortcomings in the experiment:

1. Due to experimental conditions, we have not tested more complex network architectures such as GoogLeNet, ResNet, and it is not known whether transfer learning is equally valid in these networks.

2. Because medical image data is difficult to obtain public data sets, we can't compare our models with others, and we can only verify our experimental results in a small area.

3. The accuracy of our model is high in both training set and validation set but we don ' t know if it is also high foe test set. Because we don ' t have so many samples, we can just use the cross validation set as a test set.

5.2 Prospects

Convolutional neural networks have achieved great success in the field of image recognition. However, convolutional neural networks rely on large-scale supervised learning to train and produce models with good effects. However, it is difficult to obtain enough labeled data, and the amount of unlabeled data in life is much larger than that of labeled data. We need to use transfer learning, or semi-supervised learning, unsupervised learning to overcome this limitation. Especially in the field of medical image diagnosis, we believe that the following aspects are worthy of our study:

1. The effectiveness of transfer learning: There is no theoretical research and proof of transfer learning. In other studies, transfer learning has not improved accuracy but maybe some reduced accuracy.

2. The scope of application of transfer learning: whether transfer learning can be effective in all areas, or only partially effective. In medical image diagnosis, it is effective for all images or only for a small portion of images.

3. Alternatives of transfer learning: Transfer learning can replace traditional image recognition technology. In some medical image diagnosis and treatment, some excellent traditional image recognition methods are even higher than deep convolutional neural networks and transfer learning.

Acknowledgement

Sincerely thank you my tutor, Yoshida sensei. In the two years of master's degree studying and creation of graduation master thesis, I am very fortunate to have the care and help of you! From the moment I became a graduate student, I found you with great excitement and nervousness and wanted to be your student. You are passionately teaching me to do research. I should be diligent and hard-working. I should be honest and loyal to my parents. I should be responsible and responsible in my work. When I am confused and helpless, you help me unselfishly, enlighten me, and point me to a clear direction, so that I can continue to move forward on the road of scientific research. When I have problems that cannot be solved, you take time out of your busy schedule to help me check the relevant information and answer my doubts personally, so that I can overcome my difficulties and complete my research. Whenever I have a problem and can't solve the problem, you can always patiently explain the relevant knowledge to me, and put forward some valuable suggestions and ideas to help me solve the problem quickly and efficiently and deepen my knowledge understanding.

Sincerely thank for all teachers, classmates, friends! When I was in trouble, you always encourage me and help me by all means. I had a real good time in this lab and studied the knowledge with you. I not only finished my master thesis but also learnt a lot of Japanese language, culture and had a lot of friends.

Sincerely thank for my parents! Thank you for your kindness. You encouraged me when I was disappointed and helpless. It was you who took care of me when I was sick and suffering. It was your meticulous and selfless dedication to help me finish my studies. I will reward you with good results and excellent work to reward your hard work.

Acknowledgement

Finally, sincerely thank for all those who have helped me, thank you to the experts and professors, thank you!

References

[1] Fukushima K. Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position[J]. Biological cybernetics, 1980, 36(4): 193-202.

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

[3] Hinton G E, Osindero S, Teh Y W. A Fast Learning Algorithm for Deep Belief Nets[J]. Neural Computation, 2014, 18(7):1527-1554.

[4] Krizhevsky A, Sutskever I, Hinton G E. ImageNet classification with deep convolutional neural networks[C]// International Conference on Neural Information Processing Systems. Curran Associates Inc. 2012:1097-1105.

[5] Sermanet P, Eigen D, Zhang X, et al. OverFeat: Integrated Recognition, Localization and Detection using Convolutional Networks[J]. Eprint Arxiv, 2013.

[6] Simonyan K, Zisserman A. Very Deep Convolutional Networks for Large-Scale Image Recognition[J]. Computer Science, 2014.

[7] Szegedy C, Liu W, Jia Y, et al. Going deeper with convolutions[J]. 2014:1-9.

[8] Szegedy C, Vanhoucke V, Ioffe S, et al. Rethinking the Inception Architecture for Computer Vision[C]// Computer Vision and Pattern Recognition. IEEE, 2016:2818-2826.

[9] Szegedy C, Ioffe S, Vanhoucke V, et al. Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning[J]. 2016.

[10] He K, Zhang X, Ren S, et al. Deep Residual Learning for Image Recognition[C]// Computer Vision and Pattern Recognition. IEEE, 2016:770-778.

[11] Silver D, Huang A, Maddison C J, et al. Mastering the game of Go with deep neural

networks and tree search[J]. Nature, 2016, 529(7587):484.

[12] Doi K. Computer-aided diagnosis in medical imaging: historical review, current status and future potential.[J]. Computerized Medical Imaging and Graphics, 2007, 31(4–5):198-211.

[13] Nagata R, Kawaguchi T, Miyake H. A computer-aided diagnosis system for lung nodule detection in chest radiographs using a two-stage classification method based on radial gradient and template matching[C]// International Conference on Biomedical Engineering and Informatics. IEEE, 2014:80-85.

[14] Htike Z Z, Naing W Y N, Win S L, et al. Computer-Aided Diagnosis of Pulmonary Nodules from Chest X-Rays Using Rotation Forest[C]// International Conference on Computer and Communication Engineering. IEEE, 2015:96-99.

[15] Aberle D R, Adams A M, Berg C D, et al. Reduced lung-cancer mortality with low-dose computed tomographic screening.[J]. New England Journal of Medicine, 2011, 365(5):395.

[16] Messay T, Rogers H S K. A new computationally efficient CAD system for pulmonary nodule detection in CT imagery[J]. Medical Image Analysis, 2010, 14(3):390-406.

[17] Darmanayagam S E, Harichandran K N, Cyril S R R, et al. A Novel Supervised Approach for Segmentation of Lung Parenchyma from Chest CT for Computer-Aided Diagnosis[J]. Journal of Digital Imaging, 2013, 26(3):496-509.

[18] Ashwin S, Ramesh J, Kumar S A, et al. Efficient and reliable lung nodule detection using a neural network based computer aided diagnosis system[C]// International Conference on Emerging Trends in Electrical Engineering and Energy Management. IEEE, 2012:135-142.

[19] Firmino M, Morais A H, Mendoça R M, et al. Computer-aided detection system for lung cancer in computed tomography scans: Review and future prospects[J].

Biomedical Engineering Online, 2014, 13(1):41.

[20] Yang H, Yu H, Wang G. Deep Learning for the Classification of Lung Nodules[J]. 2016.

[21] Iii S G A, Mclennan G, Bidaut L, et al. The Lung Image Database Consortium (LIDC) and Image Database Resource Initiative (IDRI): A Completed Reference Database of Lung Nodules on CT Scans[J]. Medical Physics, 2011, 38(2):915.

[22] Setio A A A, Ciompi F, Litjens G, et al. Pulmonary Nodule Detection in CT Images: False Positive Reduction Using Multi-View Convolutional Networks[J]. IEEE Transactions on Medical Imaging, 2016, 35(5):1160-1169.

[23] Xu Y, Jia Z, Ai Y, et al. Deep convolutional activation features for large scale Brain Tumor histopathology image classification and segmentation[C]// IEEE International Conference on Acoustics, Speech and Signal Processing. IEEE, 2015:947-951.

[24] Shin H C, Roth H R, Gao M, et al. Deep Convolutional Neural Networks for Computer-Aided Detection: CNN Architectures, Dataset Characteristics and Transfer Learning[J]. IEEE Transactions on Medical Imaging, 2016, 35(5):1285.

[25] Pan S J, Yang Q. A Survey on Transfer Learning[J]. IEEE Transactions on Knowledge and Data Engineering, 2010, 22(10):1345-1359.

[26] Srivastava R K, Greff K, Schmidhuber J. Highway Networks[J]. Computer Science, 2015.

[27] Glorot X, Bengio Y. Understanding the difficulty of training deep feedforward neural networks[J]. Journal of Machine Learning Research, 2010, 9:249-256.

[28] Lin M, Chen Q, Yan S. Network In Network[J]. Computer Science, 2013.

[29] Ioffe S, Szegedy C. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift[J]. 2015:448-456.