Mobile Networks and Applications (2021) 26:351–380
https://doi.org/10.1007/s11036-020-01672-7

A Review of Deep Learning on Medical Image Analysis

Jian Wang [1] · Hengde Zhu [1] · Shui-Hua Wang [1,2,3] · Yu-Dong Zhang [1,4]

Accepted: 20 October 2020 / Published online: 5 November 2020
© Springer Science+Business Media, LLC, part of Springer Nature 2020

Abstract
Compared with common deep learning methods (e.g., convolutional neural networks), transfer learning is characterized by simplicity, efficiency and its low training cost, breaking the curse of small datasets. Medical image analysis plays an indispensable role in both scientific research and clinical diagnosis. Common medical image acquisition methods include Computer Tomography (CT), Magnetic Resonance Imaging (MRI), Ultrasound (US), X-Ray, etc. Although these medical imaging methods can be applied for non-invasive qualitative and quantitative analysis of patients—compared with image datasets in other computer vision fields such like faces—medical images, especially its labeling, is still scarce and insufficient. Therefore, more and more researchers adopted transfer learning for medical image processing. In this study, after reviewing one hundred representative papers from IEEE, Elsevier, Google Scholar, Web of Science and various sources published from 2000 to 2020, a comprehensive review is presented, including (i) structure of CNN, (ii) background knowledge of transfer learning, (iii) different types of strategies performing transfer learning, (iv) application of transfer learning in various sub-fields of medical image analysis, and (v) discussion on the future prospect of transfer learning in the field of medical image analysis. Through this review paper, beginners could receive an overall and systematic knowledge of transfer learning application in medical image analysis. And policymaker of related realm will benefit from the summary of the trend of transfer learning in medical imaging field and may be encouraged to make policy positive to the future development of transfer learning in the field of medical image analysis.

Keywords Transfer learning · Medical image analysis · CT · Deep learning · MRI · Convolutional neural networks · Fine-tuning · Feature extractor · Artificial intelligence

1 Introduction

Medicine is a science that benefits all mankind, directly related to everyone's health and quality of life. As a result, medicine has always been one of the most highly regarded disciplines in the world.

Medical research is inseparable from the support of medical image analysis. Both the cutting-edge medical research conducted in the laboratory and the diagnosis made by clinicians require a large amount of evidence provided by medical image analysis to make conjecture or diagnosis. With the continuous development of medical technology, a variety of medical image means have emerged. The most widely applied medical imaging techniques include Computer Tomography (CT), Magnetic Resonance Imaging (MRI), Ultrasound (US) and X-Rays. Among these medical imaging technologies, CT has higher resolution on tissue of high density but relies on doctor's skill and exists probability of lost scans. X-Rays is convenient and low-price, suitable for first medical examination but like CT, X-Rays do harm on human bodies thus

✉ Shui-Hua Wang
shuihuawang@ieee.org

✉ Yu-Dong Zhang
yudongzhang@ieee.org

Jian Wang
jw830@le.ac.uk

Hengde Zhu
hz166@le.ac.uk

[1] School of Informatics, University of Leicester, Leicester LE1 7RH, UK

[2] School of Architecture Building and Civil engineering, Loughborough University, Loughborough LE11 3TU, UK

[3] School of Mathematics and Actuarial Science, University of Leicester, Leicester LE1 7RH, UK

[4] Department of Information Systems, Faculty of Computing and Information Technology, King Abdulaziz University, Jeddah 21589, Saudi Arabia

patients cannot take it too many times. Unlike CT and X-Rays, MRI does not have ionizing radiation and see more clearly on soft tissue, but MRI usually take a long time which some patients may not suffer, especially those wearing metal medical instruments for example cardiac pacemaker. US has advantage of detecting movements and enables doctors to watch real-time pictures inside patients' bodies. Although various medical imaging technologies have their own characteristics and doctors should choose them carefully, deep learning-based algorithms could match them well due to deep learning methods have strong robustness in image's scale and resolution.

These non-invasive imaging techniques are relatively harmless to the patient's body and allow for a qualitative and quantitative assessment of the symptoms at the site of the lesion. They are used in vital parts of the body such as brain, heart, chest, lung, kidney, liver, etc. However, artificial-based medical image analysis relies on the expertise of an experienced physician to identify the image with naked eyes. The problem of this approach is that, first of all, the number of doctors who can perform medical image analysis is extremely limited. Compared with the massive demand of medical image analysis, the number of professional doctors is far from enough in either developed or developing countries all over the world. Secondly, human eyes could be fatigued, and the level of doctors is uneven. Therefore, misjudgment often occurs, leading to wrong diagnosis and even delayed illness. Because of these defects of medical image analysis based on manual operation, scientists have been studying how to replace people through computer technology, so as to improve the efficiency and accuracy of medical image analysis.

In this context, although computer-aided diagnosis technology has a long history, it was not until the advent of convolutional neural networks that the real explosion period of automatic medical image analysis began. The convolutional layer in convolutional neural networks can extract deep features from medical images. By further processing these deep features, the convolutional neural network can perform a series of medical image analysis tasks including segmentation, detection, classification and disease prediction [1–3]. The convolutional neural network has the advantages of wide application range, high accuracy and fast analysis speed. In recent years, more and more scholars have claimed that in their experiments, the accuracy of medical image analysis system using convolutional neural network structure has surpassed that of human beings in some medical imaging data sets. But the structure of convolutional neural networks also has its shortcomings [4]. Firstly, the convolutional neural network needs to be trained before it can perform tasks. This training process is often difficult and long, requiring many skills, and the cost of the training process is relatively too high. Secondly, most algorithms based on convolutional neural network structure rely heavily on large number of labeled samples for learning. And this is in contradiction with the characteristics of medical image analysis itself. Medical images are generally difficult to collect, as well as expensive and relatively rare. Moreover, medical image labeling can only be done by professional doctors, so labeled data is even more scarce. These factors have become obstacles to the expansion of convolutional neural networks in medical image analysis.

In order to overcome these obstacles, the researchers introduced the method of transfer learning into the field of medical image analysis [5–8]. As a significant method in deep learning, transfer learning was derived from convolutional neural networks. The basic idea of transfer learning is that if a convolutional neural network is successful in solving a problem, it must have successfully learned something which could be transferred to another similar problem finding a solution faster. Compared with the common convolutional neural network, transfer learning has established its own superiorities in the field of medical image analysis. Transfer learning can take advantage of the pre-trained network as the basis without training from scratch, which greatly saves the training time and reduces the difficulty of training. In the past, the training of convolutional neural networks requires a large amount of data with labels, otherwise the model is prone to overfitting. Transfer learning only needs a small amount of labeled data to fine-tune a pre-trained network structure, which is more suitable for medical image analysis [9]. At present, in the field of medical image analysis, there are two most common strategies for applying transfer learning, fine-tuning and feature extractor. Fine-tuning is based on the pre-trained deep model, utilizing the tagged medical image data. Both the hidden layers used to extract features and the final output layers of the pre-trained networks are retrained. After fine-tuning, the network not only has stronger classification performance for the target task in the final output layer, but also has stronger and more specific feature extraction ability for the target data domain in the middle convolutional layers. The approach of feature extractor is to freeze all the other layers in pre-trained networks except for the last few layers, and then splice its own classifier reconstructing a new network model. The pre-trained networks adopted by the feature extractor strategy usually are some universal and powerful deep learning models which have been proved to be very reliable by numerous tests. It enables the newly constructed networks to earn the powerful feature extraction performance from the pre-trained network with only low training cost. In general, if you have abundant dataset of medical images labeled, you can try fine-tuning. If the labeled medical image dataset is very scarce, then the feature extractor method might be easier to avoid overfitting. Fine-tuning and feature extractor determine how many layers in the pre-trained networks participate in updating parameters during retraining process, which is essentially a trade-off problem [10]. Thus, which of the two strategies is better depends on the detailed situation.

Due to the advantages of transfer learning in the field of medical image analysis, in recent years, more and more researchers have adopted transfer learning (as the trend shown in Fig. 1) to solve problems of medical image analysis and achieved good results. These advances utilized a variety of medical imaging techniques, including CT, MRI, US, and X-Rays, covering important body parts such as brain, heart, breast, lung and kidney [11]. It should be admitted that transfer learning has been widely and deeply applied in medical image analysis. Therefore, we provided this survey paper to review, summarize and envision the development of transfer learning in medical image analysis. In the following paragraphs, we would first introduce the background knowledge of convolutional neural networks, from where transfer learning originated. Then we introduced the formal definition and categories of transfer learning. After that we reviewed the application of transfer learning in the field of medical image analysis by dividing several elemental parts of human body. Finally, we also discussed the integration of transfer learning with other deep learning technologies and the future trend, including the shortcomings, and gave conclusion on this paper.

2 Convolutional Neural Network

In recent years, the method of deep learning is more and more widely applied in the field of medical image processing. One of the most commonly used and classic methods of deep learning in medical image processing is Convolutional Neural Network (CNN). CNN not only breaks through the previous level of methods, making deep learning reach unprecedented precision in image classification, but also serves as the cornerstone of transfer learning in solving image processing [12–16]. It can be said that without the development and expansion of CNN, there will be no great prospect of transfer learning in the field of medical image processing [17, 18]. In fact, a significant amount of transfer learning approaches are based on CNN. In recent years, one of most common strategies of applying transfer learning is to utilize a classic convolutional neural network as pre-trained model, freeze some layers and then retrain a few layers by data in target domain. Another popular strategy is to cut off part of layers in pre-trained model as feature extractor then add another classifier such as Support Vector Machine (SVM). The pre-trained model plays a crucial role in transfer learning. And most of popular pre-trained models applied in transfer learning medical images analysis, such as AlexNet, VGGNet, and ResNet, use convolutional neural network structure. If we do not fully investigate convolutional neural network, it's impossible to really understand the whole aspects of modern transfer learning strategies. Therefore, it is necessary for us to understand the infrastructure of CNN, some commonly used optimization tricks, and several classic and representative CNN models that are still used on a large scale if we want to review the application of transfer learning in the field of medical images. This chapter is an introduction to this aspect.

2.1 Overall Structure

In general, when dealing with image classification tasks, CNN generally includes convolutional layers, pooling, activation and fully connected layers. The convolutional layer is mainly used to extract the features of the input image. As a downsampling operation, pooling layer is mainly used to reduce the resolution of features maps. The purpose of activation layer is to introduce nonlinear factors and improve the expression ability of neural network. The full connection layer

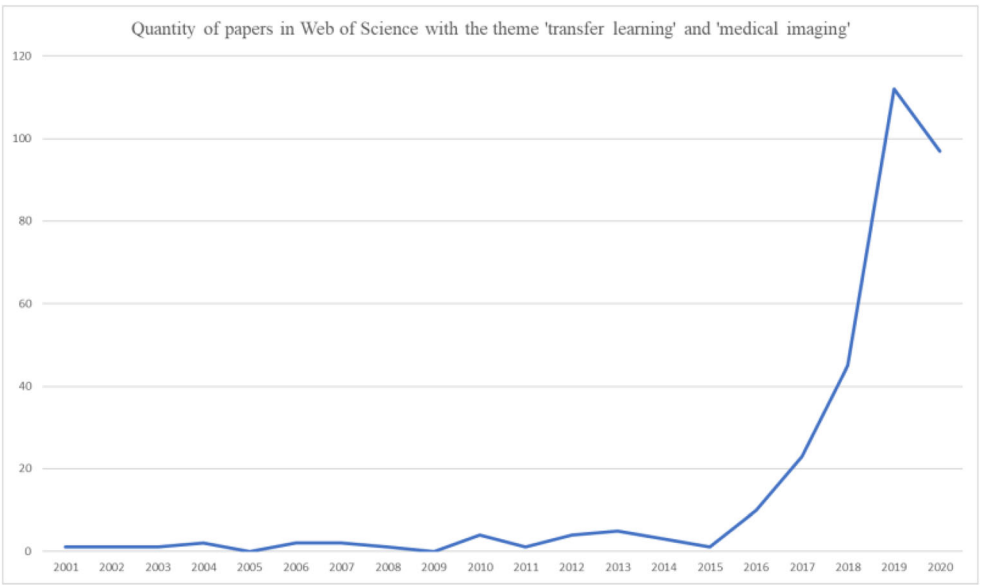

Fig. 1 Trend of papers published on transfer learning in medical image analysis

can reduce the dimension of feature maps and act as a classifier. Faced with a multi-classification problem, we usually need to use softmax as the output layer to map the probability of the result between 0 and 1. Fig. 2 shows an overall basic structure of convolutional neural networks. It is through the stacking and improvement of these basic structures that CNN continues to generate more complex and powerful neural networks.

For example, it began from VGGNet that researchers acknowledged that deeper structure brought better results. Since then researchers have been putting effort into building deeper and deeper neural network structure. As we mentioned before, convolutional layers have ability to extract features from input images. Through heaping up more and more convolutional layers, it equips neural networks with capacity of extracting deeper features from input images. During the process, the conception of block was proposed. A block usually consists of several convolutional layers with pooling layer and activation function. Blocks have become basic unit in modern convolutional neural network structure. Besides normal sequential block, more innovative block structures were proposed, including Inception and Residual Block. These blocks utilized parallel or skip connection architectures which could promote network's performance in some conditions compared to normal block and have been widely refered and applied in up-to-date research.

2.2 Convolutional Layers

In CNN, convolution layer is generally followed by the input layer to extract feature from the input layer. The convolutional kernel in the traditional convolutional layer is similar to the filter in signal processing. When the convolutional kernel slides on the image, it is only sensitive to the image with a specific feature. Therefore, different features can be extracted from the image of the input layer through different convolutional layers. Generally, CNN contains multiple convolutional layers. The former convolutional layer extracts some basic features, while the latter convolutional layer extracts advanced features from the basic features. For example, if we need to judge a cat, the features extracted from the first convolutional layer may be only some edges or lines, while the features extracted from the second convolutional layer may be some local organs, such as the cat's eyes, nose and ears. Through this layer-by-layer approach, CNN finally learned it was a cat that matched these features. This is the usual standard convolutional layer structure.

In order to discuss convolution structure conveniently, we first give brief definition to most essential parameters of convolution then investigating several different kinds of convolutions and their traits. Let us define k as kernel size, t as input image's size, p as zero-padding, s as stride and u as output feature map's size. Kernel is referred to feature extractor, usually represented by a $k \times k$ matrix. To make problem easy to explain, we normally suppose that the input image is of $t \times t$ pixels and the output image is of $u \times u$ pixels. Padding is often utilized to supplement extra pixels of value zero around the input image to ensure convolutional neural networks reaching deeply as we need. The ordinary operation of convolution is to perform kernel on input image one by one step, in which case the stride s equals to one. But things would be different if we acquire other performance and the stride could be other values.

2.2.1 Standard Convolution

We investigate the simplest and most classic convolution structure. In this condition, the output feature map's size u could be written as:

$$u = t - k + 1 \tag{1}$$

According to Fig. 3, in this simplest example, the convolution does not have padding and set stride to one. That is to

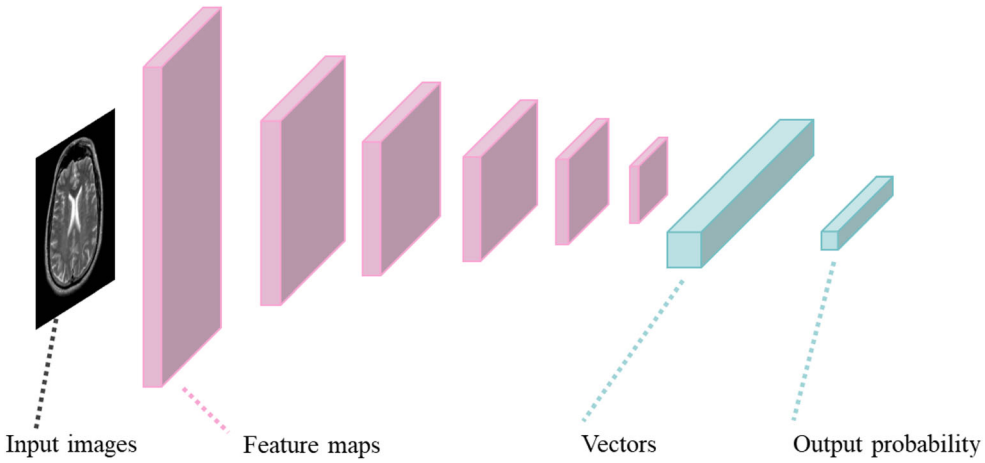

Fig. 2 Basic structure of typical convolutional neural networks

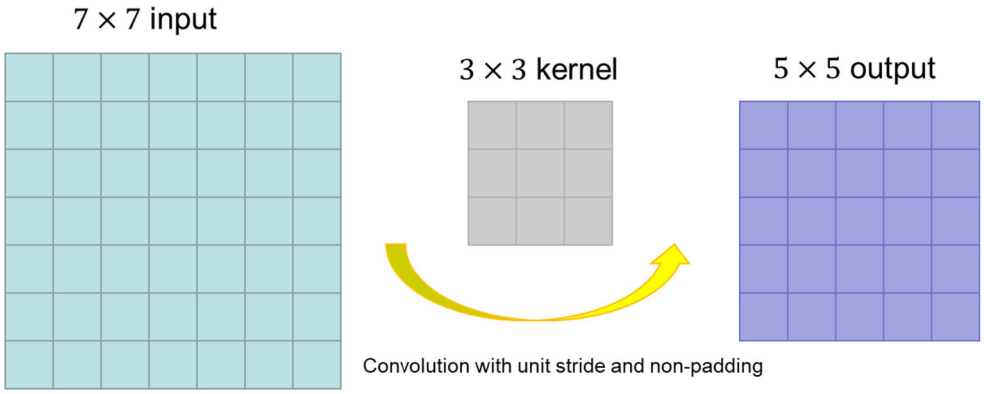

Fig. 3 A simple example of convolution

say, $t = 7$, $k = 3$, thus $u = 5$. We consider a more complex situation that we add a circle of zero-padding around the input image. Then the equation should be recorded in Eq. (2).

$$u = t - k + 2p + 1 \tag{2}$$

According to Fig. 4, in this more complex situation, the convolution has one zero-padding and still keep stride as one. That is to say,

$$\begin{aligned} t &= 7 \\ k &= 3 \\ p &= 1 \\ u &= 7 \end{aligned} \tag{3}$$

Moreover, we could observe one of most important advantage of zero-padding, to get output as same size as input image's without compressing pixels. Without zero-padding, it is easily to imagine that after convolutional operations layer by layer, the output feature map's size gets smaller and smaller. As a result, we could not apply deep convolutional architecture in this circumstance. But with zero-padding, we keep output with the same size of input image, which equips us ability to design deep convolutional neural networks.

2.2.2 Strided Convolution

With the continuous development of CNN, more and more new convolutional layer structures have been proposed. Strided convolution is based on standard convolution, and the strided convolutional kernel is shifted by more than 2 pixels at a time. Using the definition above, the equation should be summarized as:

$$u = \left\lfloor \frac{t - k + 2P}{s} \right\rfloor + 1 \tag{4}$$

According to Fig. 5, the 3×3 filter stride three pixels every move for next calculating. In this example, $t = 7$, $k = 3$, $p = 1$, $s = 3$, thus $u = 3$. In this way, we can obtain smaller feature maps and achieve the effect similar to pooling to some extent.

There is another convolution aiming to reducing quantity of parameters called grouped convolution. It was first proposed in AlexNet. Grouped convolution divides convolutional kernels into several groups by means of neural network segmentation. The feature maps obtained are only part of the original one, which can be processed in parallel with multiple GPUs.

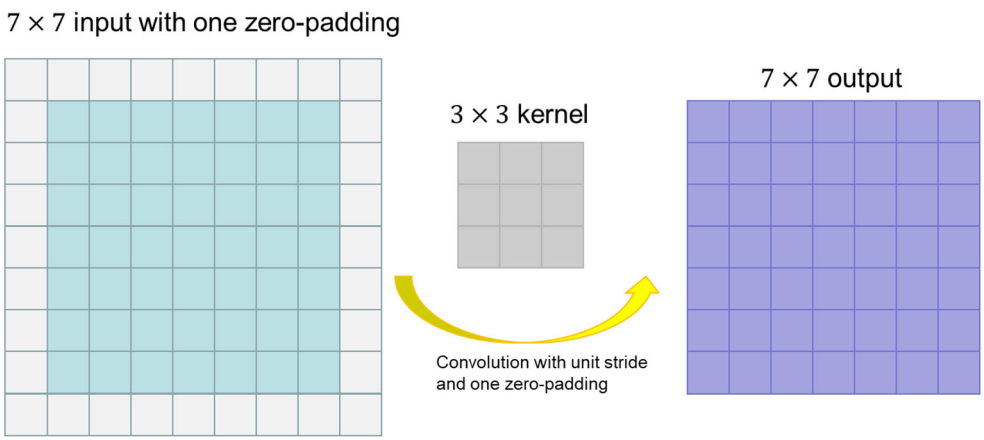

Fig. 4 Convolution with zero-padding

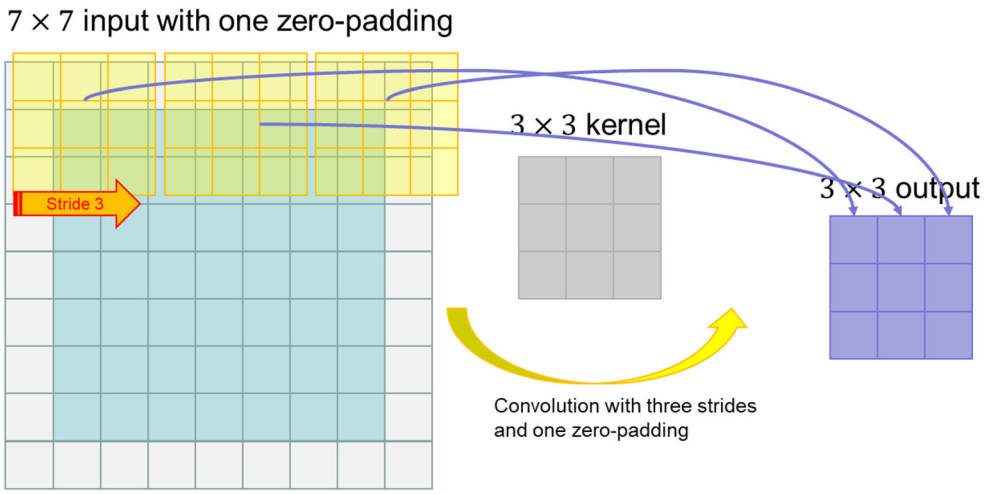

Fig. 5 Strided convolution with zero-padding

2.2.3 Grouped Convolution

Figure 6 demonstrates how grouped convolution could decrease quantity of parameters. Let us first analyze standard convolution. The size of input feature map is

$$SizeInput = E \times F \times m \tag{5}$$

and the size of filter is

$$SizeFilter = e \times f \times m \tag{6}$$

And in normal convolution operation, we need n filters to achieve output feature map which size is

$$SizeOutput = E \times F \times n \tag{7}$$

In this progress, we need to figure out

$$e \times f \times m \times n = Q \tag{8}$$

parameters.

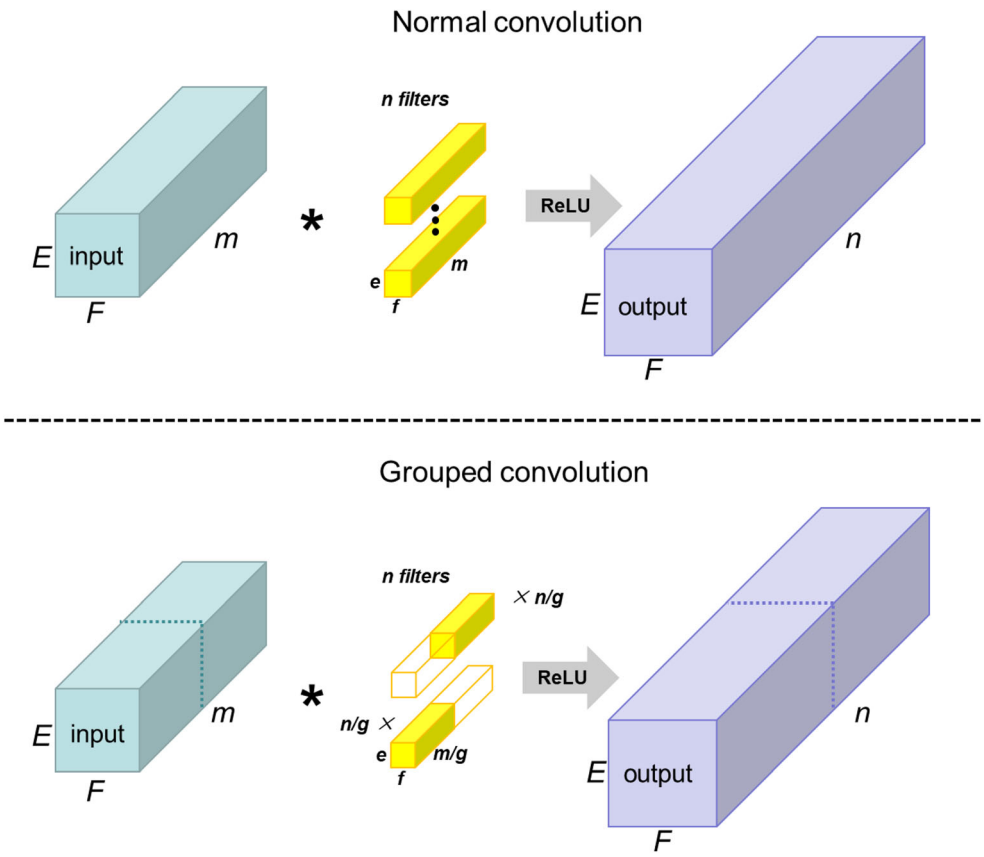

Fig. 6 Comparison between normal convolution and grouped convolution

While in grouped convolution, we divided input feature map into g groups by m channels. Each group of input feature map is of

$$SizeGroupInput = E \times F \times \frac{m}{g} \tag{9}$$

with its corresponding filter of size

$$SizeGroupFilter = e \times f \times \frac{m}{g} \tag{10}$$

getting result of grouped output feature map which size is

$$SizeGroupFeatureMap = E \times F \times \frac{n}{g} \tag{11}$$

After concreting g groups of grouped output feature map, we finally get ultimate output feature map which size is

$$SizeOutput = E \times F \times n \tag{12}$$

It is easily to find that we only use

$$e \times f \times \frac{m}{g} \times \frac{n}{g} \times g = \frac{e \times f \times m \times n}{g} = \frac{Q}{g} \tag{13}$$

parameters which is largely less than parameter used in normal convolution operation. Taking another aspect to think about this problem, during normal convolution, every point in output feature map is produced by filter which size is

$$SizeFilter = e \times f \times m \tag{14}$$

while during grouped convolution every point in output feature map is produced by filter which size is only

$$SizeGroupFilter = e \times f \times \frac{m}{g} \tag{15}$$

That is why grouped convolution could reduce quantity of parameters.

2.2.4 Dilated Convolution

Last, we introduce dilated convolution. The size of convolutional kernel in dilated convolution is no longer corresponding to the pixel in the input image but corresponding to a larger size of the input image for convolution operation. The advantage of dilated convolution is to make convolution have larger receptive field with low extra computational cost. In dilated convolution, a key hyperparameter written as d is proposed to represent dilation rate. For $d = 1$, that means normal convolution. For $d = 2$, that means there exists $d - 1$ in this case one additional space between every point where kernel performs convolution operation. We could observe its corresponding relationship in Fig. 7. In dilated convolution, the corresponding area of kernel is enlarged to a wider range. Therefore, we could define the size of kernel's actual corresponding area as k', and we give the equation of k':

$$k' = k + (k-1)(d-1) \tag{16}$$

We applied k' instead of k in Eq. (3) getting:

$$u = \left\lfloor \frac{t-k-(k-1)(d-1)+2P}{s} \right\rfloor + 1 \tag{17}$$

This is the equation refers to output size in dilated convolution. In the example of Fig. 7, $t = 7$, $k = 3$, $d = 2$, $p = 0$, $s = 1$, thus $u = 3$.

In the end, we summarized these various convolutions as following Table 1.

2.3 Pooling

We know that the convolutional layer extracts enough features from the input image. But in many cases, too many features are not always a good thing. The extracted features may contain information that we do not care much about, and this redundant information can make the entire network slow and bloated, so we want to remove the redundancy. Pooling layer is designed to carry out down-sampling operation on extracted feature maps, compress the resolution of feature maps, and only retain important feature information. Pooling layer is also a convolution operation mathematically. Unlike the convolution kernel of convolutional layer, the parameters of pooling layer are usually fixed. The advantage of pooling layer is that, firstly, it has translation invariance. Secondly, parameters of pooling layer are fixed, so the quantity of parameters in entire neural networks can be reduced.

In traditional CNN, there are usually two methods: max pooling and average pooling. Max pooling is to select the maximum value from a local domain of the image as the representative, which can better preserve the texture features of the image. Average pooling uses selection of the average value as the representative from a local domain of the image, which can better preserve the features of the overall image's data. Both max pooling and average pooling could be regarded as convolution which stride the same quantity pixels as its kernel's. But another pooling method strides less pixels than its convolution kernel's size, called overlapping pooling. It's easy to understand that overlapping pooling could store more information in feature maps compared to max pooling and average pooling. What's more, scientists proposed spatial pyramid pooling which adopts different scale pooling kernels and strides. With spatial pyramid pooling, feature maps of different sizes could be handled. And because of spatial pyramid pooling using different sclaes of pooling kernel then converging the results, it helps promote network structure's accuracy and robustness.

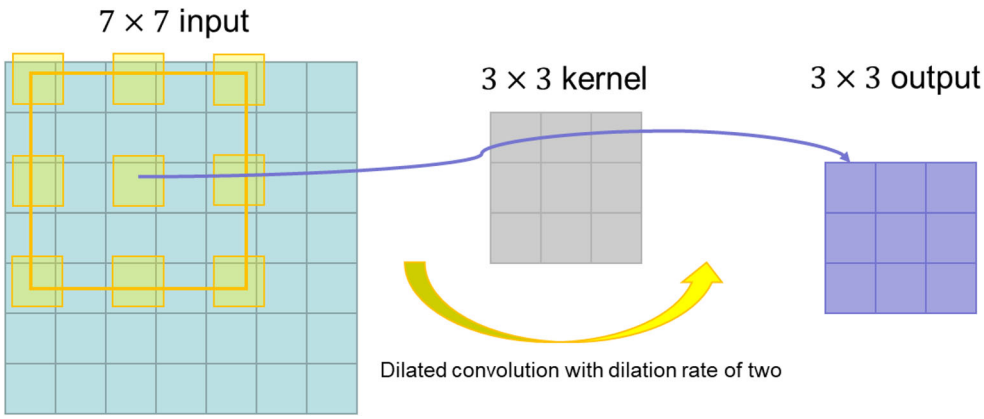

Fig. 7 Dilated convolution

Last, we provided Table 2 to compare these variety of pooling methods.

2.4 Activation

The function of activation is to introduce nonlinearity into CNN. In a practical problem, the data is often not separable linearly. Without activation, it is difficult for CNN to achieve a good effect on linearly indivisible data. Sigmoid and Tanh are two of earliest proposed activation functions.

The equation of Sigmoid could be written as:

$$sigmoid(x) = \frac{1}{1 + e^{-x}} \tag{18}$$

The equation of Tanh could be written as:

$$tanh(x) = \frac{e^{x} - e^{-x}}{e^{x} + e^{-x}} \tag{19}$$

At present, in the field of image and vision, the most commonly used activation function is ReLU. Compared with Sigmoid or Tanh, ReLU can converge more quickly and effectively alleviate the problem of gradient vanishing. The equation of ReLU could be written as:

$$ReLU(x) = \begin{cases} x, x > 0 \\ 0, x \leq 0 \end{cases} \tag{20}$$

On the basis of ReLU, a series of improved activation functions are derived as well. Leaky ReLU, compared to ReLU, when $x < 0$, y did not equal to 0 but a very small negative, letting the line of function continue declining at a small gradient. The equation of Leaky ReLU could be written as:

$$LReLU(x) = \begin{cases} x, x > 0 \\ 0.01x, x \leq 0 \end{cases} \tag{21}$$

In practice, the slope of the $x < 0$ part of Leaky ReLU might not be easy to determine, so PReLU was designed. PReLU can adaptively learn the slope parameters for the part of $x < 0$ from the data. The equation of PReLU could be represented as:

$$PReLU(x) = \begin{cases} x, x > 0 \\ ax, x \leq 0 \end{cases} \tag{22}$$

where a is a very small value not fixed and determined by other parameters in specific neural networks.

Another way is to take the slope of $x < 0$ as a random parameter to sample from a given range, which is called Randomized ReLU (RReLU). The difference between RReLU and PReLU (shown in Fig. 8) is that the slope parameter a is not fixed during every training process but changed into a random value of the given range when next epoch of training begins. The equation of RReLU could be recorded as:

Table 1 Variety of convolution

Convolution	Zero-padding	Stride	Groups	Dilation rate	Benefits
Normal convolution	0	1	0	1	Basic and Simple
Convolution with padding	Usually more than 1	1	1	1	Ensure networks reach deep
Strided convolution	Flexible	Usually more than 2	1	1	Like pooling
Grouped convolution	Flexible	Flexible	Usually more than 2	1	Reduce quantity of parameters
Dilated convolution	Flexible	Flexible	1	Usually more than 2	Expand receptive field

Table 2 Comparison of pooling methods

Pooling	Size	Stride	Strategy	Benefits
Max pooling	Fixed	= Size	Fetch max pixel of local region	Preserve texture feature
Average pooling	Fixed	= Size	Calculate mean of local pixels	Preserve background information
Overlapping pooling	Fixed	< Size	Usually fetch max pixel of local region	Better representative ability
Spatial pyramid pooling	Flexible	= Size	Usually fetch max pixel of local region	Overcome various scales and higher accuracy

$$RReLU(x) = \begin{cases} x, x > 0 \\ a\,x, x \leq 0 \end{cases} \tag{23}$$

where $a \sim U(l, u)$, $l < u$ *and* $l, u \in [0, 1)$

Besides family of ReLUs, exponential linear units (ELU) was proposed to make the average value of activation function closer to zero by introducing exponential equation instead of linear equation when $x < 0$. Due to ELU's mean value closer to zero, it is believed that ELU has faster learning speed than ReLUs in some conditions. The equation of ELU could be noted as:

$$ELU(x) = \begin{cases} x, x > 0 \\ \beta(e^x - 1), x \leq 0 \end{cases} \tag{24}$$

And if adding a scale parameter λ infront of ELU, we could get scaled exponential linear units (SELU). The equation of SELU could be written as:

$$SELU(x) = \lambda \begin{cases} x, x > 0 \\ \beta(e^x - 1), x \leq 0 \end{cases} \tag{25}$$

In short, it depends which activation function should be applied in convolutional neural networks. We need to choose appropriate ReLU function (summarized in Table 3) according to the actual problems to be solved.

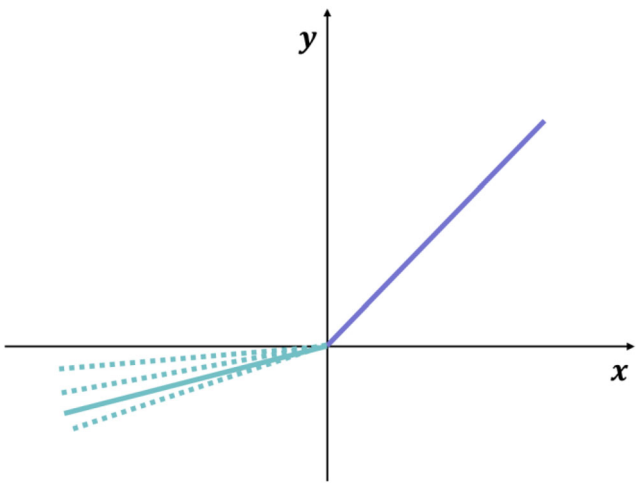

Fig. 8 Randomized ReLU

2.5 Fully Connected Layer

Fully connected layer plays the role of "classifier" in the whole convolutional neural network. If operations such as convolutional layer, pooling layer and activation function layer map the original data to the hidden feature space, fully connected layer maps the learned 'distributed feature representation' to the sample space. Fully connected layer is also achieved by convolution operation, based on feature maps after convolutional layer, pooling and activation function. Fully connected layer uses a corresponding convolution kernels executing convolution on feature maps to get a one-dimensional vector. The purpose is weighting all the characteristics from neural networks and at the same time reducing spatial dimension of these characteristics, making it convenient for following softmax layer to output classification probability.

In traditional CNN, usually more than one fully connected layers are used to construct fully connected network. Parameters of fully connected layers take up a large proportion of the total Convolutional Neural Network. Too many parameters in fully connected layer tend to make the model appear bloated and lead to overfitting. Fully connected layer is not irreplaceable. Global average pooling can directly apply average pooling operation on the whole feature space, and directly output one-dimensional vector as the result, greatly reducing the number of parameters in the model. However, global average pooling is not always superior to fully connected layer, especially in transfer learning. Because most of parameters are included in fully connected layer, which has more room for adjustment. So, models with fully connected layers often perform better in transfer learning than those without fully connected layers.

3 Advanced Techniques of CNN

Since convolutional neural networks faced the world, scientists have been always searching new methods and exploring novel techniques to improve CNN. In this section, we introduced some of most popular and useful techniques that have been widely applied in modern CNN architecture. In general, these advanced techniques of CNN aim to elevate the

Table 3 Various activation function

Activation	Saturability	Symmetry about the origin	Speed of convergence	Output range	Characteristics
Sigmoid	Saturated	No	low	(0, 1)	Gradient vanishing
Tanh	Saturated	Symmetrical	Relatively low	(−1, 1)	Mean value equals to zero, faster than sigmoid but still with gradient vanishing
ReLU	No	No	Fast	[0, +∞)	Relieve gradient vanishing but with dead neurons and excursion
Leaky ReLU	No	No	Restricted fast	(−∞, +∞)	Mitigate dead neurons
PReLU	No	No	Relatively fast	(−∞, +∞)	Faster than Leaky ReLU
RReLU	No	No	Relatively fast	(−∞, +∞)	More flexible
ELU	No	No	Fast	*usually*(−1, +∞)	Faster than ReLU
SELU	No	No	Fast	*usually*(−1, +∞)	More flexible

accuracy of neural networks and make training process easier and faster. Also, we discussed some classic pre-trained models. These advanced techniques and classic pre-trained models have been the backbone in modern transfer learning technology which we should pay attention to.

3.1 Batch Normalization

Because practical problems are challenging, there is a tendency to use deeper and deeper network structure. Deep neural network tuning is very difficult and often causes internal covariate shift. Internal covariate shift refers to the phenomenon that when parameters change in the network, the data distribution of internal nodes also changes. There are two main problems brought by it. One is that the upper network needs to adjust constantly to adapt to the change of input data distribution, which slows down the learning speed of the network. The second is to make the activation function easily fall into the gradient saturation zone and reduce the speed of network convergence.

A very efficient way is to use batch normalization, an operation that normalizes the output signal into an ideal range. Given the input of a batch belonging to one layer of neural networks is

$$X = \left[x_1, x_2, x_{3,\cdots}, x_{\mathrm{n}}\right] \tag{26}$$

in which x_i means a sample and n means batch size.

First, calculate the mean of the elements from the mini-batch,

$$\varphi_B = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{27}$$

Second, calculate the variance the mini-batch,

$$\omega_B^2 = \frac{1}{n}\sum_{i=1}^{n} \left(x_i - \varphi_B\right)^2 \tag{28}$$

Then, we could perform normalization on each element from the mini-batch,

$$x_i^{'} = \frac{x_i - \varphi_B}{\sqrt{\omega_B^2 + \varepsilon}} \tag{29}$$

Last step, to compensate for the non-linear expression of the network, we need to scale and shift the original output,

$$y_i = \alpha_i \cdot x_i^{'} + \beta_i \tag{30}$$

The advantage of batch normalization lies in that, firstly, the data input into each layer of network is within a certain range through normalization. Thus, the latter layer network does not need to constantly accommodate the changes of input in the underlying layer of network, realizing the decoupling between layers of network, which is conducive to improve speed of learning the whole neural network. Secondly, batch normalization makes the model less sensitive to parameters in the network, increasing the network's adaptability to parameter's range and making network's learning more stable. At the same time, batch normalization can alleviate the problem of gradient vanishing because it can suppress the impact of changes in the underlying network accumulating into the upper network and avoid activation function from falling into the gradient saturation zone during training. Finally, batch normalization adds random noise to the network's learning process, which bring regularization effect on some level.

3.2 Dropout

In deep learning, we often encounter overfitting. Especially when our training data samples are relatively small, while the network adopted is relatively complex and the parameters are relatively large, it is easy to lead to overfitting, and only a very bad model can be obtained. So, we usually use dropout when training deep neural networks. During each training, we randomly set half of the nodes in the hidden layer of the neural

network to 0, which is equivalent to randomly ignoring half of the feature detectors, and the overfitting problem can be significantly alleviated. The contribution of dropout is to weaken the interaction between nodes in the hidden layer of neural network, punish some neurons that are too prominent, and reduce the dependence of the whole network on these prominent neurons. Therefore, dropout has become a common means to solve overfitting problem.

3.3 Regularization

To solve the problem of overfitting, in addition to dropout, another common method is regularization. Generally, overfitting occurs due to some parameters of the nodes in hidden layer are over-trained, so that these parameters can have a great impact on the prediction results of the whole model. As a result, the network is very close to truth within the training data but has a large error within the testing data. In other words, the whole network is so dependent on some part of parameters in hidden layers that it is almost kidnapped by these parameters. The idea of Regularization is to punish these parameters which are prone to being over-trained by adding the influence factors of hidden layer parameters' distribution into loss function. So that the previously over-dependent parameters can be suppressed in training, which can effectively alleviate the problem of overfitting.

3.4 Weight Initialization

3.4.1 Zeros and Constant

The simplest initialization method is to initialize all the weight parameters to 0 or a constant but using this method will cause all the neurons in the network to learn the same characteristics. The reason is that no matter how many iterations of feed-forward propagation and backpropagation are performed, the weight values between any two connected hidden layers remain the same and symmetric. We originally expect that different neurons could learn different parameters, but because the parameters are the same, different neurons could not learn different features at all. Every layer seems to contain only one neuron. Therefore, it is necessary to initialize weight values randomly.

3.4.2 Random Normal

Initializing weight values randomly following random normal distribution has two potential issues: vanishing gradients or exploding gradients. When initializing weights to a small random number, the model can run well for a period of time, but with the increase of time, the gradient starts to approach to zero in propagation, which will lead to the vanishing gradients and slow learning. When initializing weights to a large random number, it can lead to exploding gradient problem during training.

3.4.3 Random Uniform

Random uniform initialization draws weight values randomly from a uniform distribution given lower and upper bound of the range of random values. Every number within range has equal probability to be picked. Its probability density function at the two boundaries a and b is given by,

$$f(w) = \begin{cases} \frac{w-a}{b-a} & for\ a \le w \le b \\ 0 & for\ w < a\ or\ w > b \end{cases} \tag{31}$$

The cumulative distribution function is given by,

$$F(w) = \begin{cases} 0, & for\ w < a \\ \frac{w-a}{b-a}, & for\ a \le w \le b \\ 1, & for\ w > b \end{cases} \tag{32}$$

3.4.4 Truncated Normal

Truncated normal initialization is similar to random normal initialization. The difference is that values more than two standard deviations from the mean would be discarded and reassigned. The benefit of using truncated normal distribution is to prevent saturation of neurons. For example, if we use sigmoid as the activation function, once the input of activation function is too small or too large, it may cause activation value to be too small or too large, and thus enter the saturation zone. Once in the saturated zone, these neurons die and never renew. Weight values from truncated normal distribution derive from a normal distribution with mean μ and variance σ^2 and lie within the interval (a, b), with

$$a = \mu - 2\sigma \tag{33}$$

$$b = \mu + 2\sigma \tag{34}$$

Its probability density function f is given by

$$f(w; \mu, \sigma, a, b) = \begin{cases} \frac{1}{\sigma} \frac{\phi\left(\frac{w-\mu}{\sigma}\right)}{\Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right)} & for\ a \le w \le b \\ 0 & for\ w < a\ or\ w > b \end{cases} \tag{35}$$

Here

$$\phi(\xi) = \frac{1}{\sqrt{2\pi}} exp\left(-\frac{1}{2}\xi^2\right) \tag{36}$$

is the probability density function of the standard normal distribution and

$$\Phi(x) = \frac{1}{2}\left[1 + erf\left(\frac{x}{\sqrt{2}}\right)\right] \tag{37}$$

is its cumulative distribution function.

The error function, donated by $erf(x)$, is defined by

$$erf(x) = \frac{1}{\pi}\int_{-x}^{x} e^{-t^2} dt \tag{38}$$

3.4.5 Orthogonal

Initializing weights with orthogonal matrix is beneficial to propagation of gradients in deep nonlinear networks. Orthogonal matrixes are norm-preserving, which keeps the norm of input constant throughout the network. Therefore, it helps with the problem of exploding gradients and vanishing gradients. Another property of orthogonal matrix that columns are orthonormal to one another help weights to learn different input features [19].

3.4.6 Identity

Weight values are initialized with identity matrixes as a square tensor with 0's everywhere except for 1's along the diagonal. In practice, multiplicative factor can be applied to the identity matrix. This initialization method is only used to generate 2D square tensors. Compared to zero and constant initialization, identity weight tensors break the symmetry by adding 1's at the diagonal, which can help improve performance. However, when each layer of the network is activated by a linear function, the activation values will either decrease or increase exponentially with layers, leading to either vanishing gradients or exploding gradients.

3.4.7 Xavier Initialization

The core idea of Xavier Normal Initialization is that, to keep information flowing efficiently in forward-propagation, the deviations of every two connected layers' output should be the same [20]. Xavier's deduction based on several hypotheses before: 1) using symmetric activation function with unit derivation at 0; 2) initializing weights independently; 3) the same input features variances; 4) in a linear regime at the initialization. From the property of uniform distribution variance in probability statistic, the final initialization distribution of Xavier can be obtained as follows, with n_i the size of layer i,

$$\boldsymbol{W \sim U}\left[-\frac{\sqrt{\boldsymbol{6}}}{\sqrt{\boldsymbol{n_i + n_{i+1}}}}, \frac{\sqrt{\boldsymbol{6}}}{\sqrt{\boldsymbol{n_i + n_{i+1}}}}\right] \tag{39}$$

Xavier initialization following normal distribution draws samples from a truncated normal distribution centered on 0 with standard deviation = sqrt(2 / (n_i + n_{i+1})) where n_i is the number of input units in the weight tensor and n_{i+1} is the number of output units in the weight tensor.

$$\boldsymbol{stddev} = \sqrt{\frac{\boldsymbol{2}}{\boldsymbol{n_i + n_{i+1}}}} \tag{40}$$

3.4.8 He Initialization

Xavier initialization is based on the hypothesis that the model is using a linear activation function. This assumption is not valid for the ReLU activation function. With the increasing depth of network using ReLU as activation function, the network with initialized weight values following simple normal distribution and Xavier initialization has difficulty to converge. He et al. proposed a novel initialization method that works well with ReLU [21]. Compared to model using Xavier initialization, using He initialization increases the rate of convergence, although there is no clear superiority on accuracy between two models. In this method, weight values follow a zero-mean normal distribution with standard deviation as follows, with n_i the size of layer i,

$$stddev = \sqrt{\frac{2}{n_i}} \tag{41}$$

Similarly, He initialization following uniform distribution draws weight values from a uniform distribution as follows, only considering the size of input layer, with n_i the size of layer i,

$$W \sim U\left[-\frac{\sqrt{6}}{\sqrt{n_i}}, \frac{\sqrt{6}}{\sqrt{n_i}}\right] \tag{42}$$

3.4.9 Lecun Initialization

To prevent back-propagated gradients from vanishing or exploding so that learning can proceed and allow the network to learn the linear part of the mapping, it is important make weights range over the sigmoid's linear region. Lecun, et al. achieved this by normalizing training set and requiring that every layer has a constant variance of the activations $\sigma = 1$ [22]. To ensure a standard deviation of approximately 1 at the output of each layer, weights are set to values randomly chosen from a distribution with mean zero and standard deviation as follows, with n_i the size of layer i,

$$stddev = \sqrt{\frac{1}{n_i}} \tag{43}$$

Lecun initialization following uniform distribution draws weight values from a uniform distribution as follows, with n_i the size of layer i,

$$W \sim U\left[-\frac{\sqrt{3}}{\sqrt{n_i}}, \frac{\sqrt{3}}{\sqrt{n_i}}\right] \tag{44}$$

3.4.10 Positive Unitball Initialization

In this method, the sum of weight values of each layer is set to 1. This can be implemented by assigning values from a uniform distribution in [0, 1] and dividing these initialized values by the sum of them. This method can avoid initial weight values being too large to enter the saturation zone of activation functions, such as sigmoid function.

Table 4 shows the summary of different weight initialization methods. For Xavier Initialization, He Initialization, Lecun Initialization and Positive Unitball Initialization, weight values could be drawn from either a normal distribution or a uniform distribution.

4 Transfer Learning

In the field of medical image, one situation is often encountered. Database is very difficult and expensive to establish, so that the sample data is usually scarce. And the other situation is that we want to learn something new from a problem we solved in the past and quickly move on to the next task. That is why we need transfer learning. In general, we call existing knowledge source domain and call new knowledge to be learned target domain.

To give formal definition of transfer learning, we need to define two underlying concepts first. A domain, written as D, is defined to have two aspects, feature space X and marginal distribution $P(X)$. In this condition, we could represent the domain as

$$D = \{X, P(X)\} \tag{45}$$

With the definition of domain, we could define a task T as

$$T = \{\gamma, P(Y|X)\} \tag{46}$$

in which γ represents label space and the function $P(Y|X)$ predicts corresponding label based on feature space and could also be written as function η. Thus, we get the definition of a task,

$$T = \{\gamma, \eta\} \tag{47}$$

Now, we introduce the definition of transfer learning. There exists a source domain D_s with its matching task T_s while there also exists a target domain D_T with its relevant task T_T. If $D_s \neq D_T$ or $T_s \neq T_T$, transfer learning is a process which aims to learn the target probability distribution prediction function η_T in D_T using the knowledge learnt from D_s and T_s.

Transfer learning is an approach of how to transfer knowledge from source domain to target domain (seeing Fig. 9). There are four most commonly used methods of transfer learning: instance based transfer learning, feature based transfer learning, parameter based transfer learning, and relation based transfer learning.

4.1 Instance Based Transfer Learning

Instance based transfer learning is simple and easy to implement. What we need to do is comparing the source domain and target domain, marking the data that is similar to those in target domain of source domain, and increasing the weight of

Table 4 Various weight initialization methods

Initialization	Normal distribution	Uniform distribution	Random initialization	Characteristics
Zeros and Constant	No	No	No	Different neurons learn the same features
Random Normal	Yes	No	Yes	Potential issues: vanishing gradients or exploding gradients
Random Uniform	No	Yes	Yes	Every value within range has equal probability to be picked
Truncated Normal	Yes	No	Yes	Prevent saturation of neurons
Orthogonal	No	No	Yes	Beneficial to propagation of gradients
Identity	No	No	No	Break the symmetry of Zeros Initialization
Xavier Initialization	Could be	Could be	Yes	Keep information flowing efficiently in forward-propagation
He Initialization	Could be	Could be	Yes	Works well with ReLU
Lecun Initialization	Could be	Could be	Yes	Prevent back-propagated gradients from vanishing or exploding
Positive Unitball Initialization	Could be	Could be	Yes	Prevent saturation of neurons

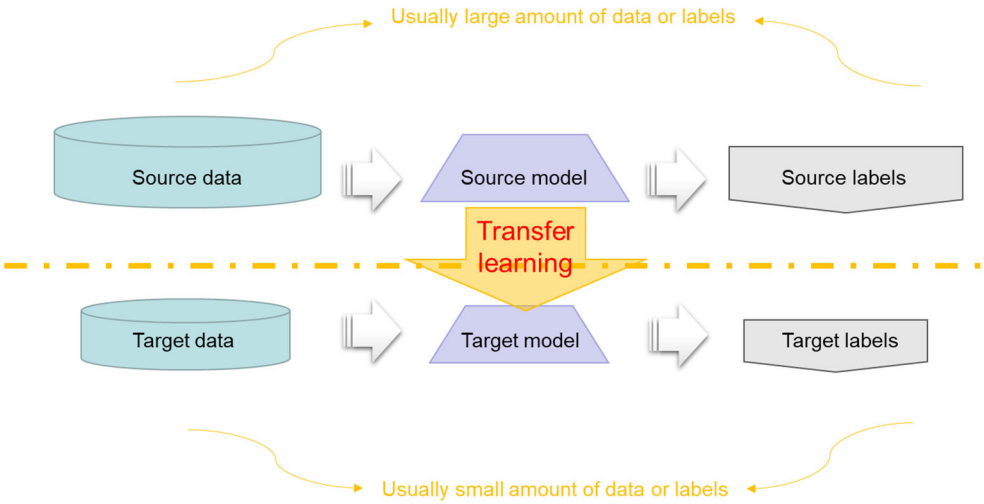

Fig. 9 How transfer learning works from source domain to target domain

this part of data. This operation is equivalent to extracting the part of data closest to those of target domain's from source domain then to match target domain. The disadvantage of this approach is that it is unstable, more empirical, and does not always exist a subset of data in source domain that happens to be very close to target domain's.

4.2 Feature Based Transfer Learning

Feature based transfer learning is firstly based on the assumption that target domain and source domain share some overlapping characteristics in common. Then we can transform source domain and target domain into a same space through feature transformation. When source domain and target domain are in the same space, data in source domain and target domain will have a similar distribution as well. So, we can use machine learning to solve the rest of work. The strong point of feature based transfer learning is that it works well while the its weak point is that it is often difficult to calculate.

4.3 Parameter Based Transfer Learning

Parameter based transfer learning is based on another assumption that source domain and target domain share part of the model's parameters, provided that source problem and target problem have some correlation. Let us say we have a CNN model that has been trained to tell difference between cats and dogs, and now we are going to move on to the new problem of distinguishing different species of cats. So, we think that the model for distinguishing between cats and dogs should have the ability to learn some of the basic characteristics of cats which can be used in new models. There are usually two specific approaches. One is to initialize the new model with the parameters of source model and then fine-tune it. Second, we solidify source model or part of layers in source model as feature extractors in the new model and then add the output layer for target problems to learn from this basis, which can effectively use the previous knowledge and reduce the cost of training. These two are the most popular transfer learning methods under the current trend of deep neural networks.

4.4 Relation Based Transfer Learning

Relation based transfer learning requires the assumption that source and target domains are similar so that this time they share some kind of logical relationship. Attempts to transfer logical relationships from source domain to target domain constitute the core idea of relation based transfer learning.

Finally, we provide a brief summary and comparison between these four methods of applying transfer learning through Table 5.

4.5 Classic Pre-Trained Models

This section mainly introduces several classic pre-trained models. On the one hand, these models were sensational at that time and played a significant role in promoting the development of deep neural network. On the other hand, many of them are still widely used, especially as the prototype and foundation of transfer models in transfer learning. Pre-trained models are applied in transfer learning mainly in two ways, fine-tuning and feature extractor. Fine-tuning adopts pre-trained models but re-trains them with target datasets only on last few layers of pre-trained networks, keeping parameters of former layers fixed away from weights updating. Feature extractor usually uses pre-trained models except for their fully-connected layers extracting deep features from images for further process. It is obvious that either fine-tuning or feature extractor demands pre-trained models as basis. Even we could say that in transfer learning better pre-trained models almost mean better performance. Therefore, only by mastering these classical pre-trained convolutional neural networks can we truly understand the development of transfer learning.

Table 5 Four methods to apply transfer learning

Transfer what	Process	Characteristics
Instance	Use target data within instance of source domain	Suitable for situation that cannot re-using data of source domain directly
Feature representation	Apply feature representation (e.g. feature extractor in CNN) of source domain into target domain	Narrow the gap between source domain and target domain but usually rely on labeled data
Parameter	Adopt parameters in source domain as initialization and give extra weight to supplement loss in target domain during re-training (e.g. fine-tuning)	Compared to training from scratch, it is easy to train new neural networks with fast speed based on pre-trained model
Relational-knowledge	Learn the relationship within data points of source domain	Compatible for data with dependency and identical distribution

4.5.1 AlexNet

As it is shown by Fig. 10, AlexNet adopts three sizes of convolutional kernel, 11×11, 5×5 and 3×3. Features are continuously extracted via convolution and the size of feature map is also concentrated with maxpooling method. Finally, the final classification result is outputted through the combination of two fully connected layers and one softmax layer. Other highlights in AlexNet include the use of ReLU as activation function, the use of overlapping pooling in the first layer, and the introduction of dropout method in fully connected layers, which laid a good foundation for the subsequent CNNs.

4.5.2 VGGNet

VGGNet shows the superiority of deep network's structure. VGG (showing in Fig. 11) utilizes smaller convolutional kernel than AlexNet, only 3×3, but the network deepens to 16 even 19 layers. Starting from VGGNet, researchers gradually moved toward the idea of small convolutional kernel but with layers deeper and deeper to construct convolutional neural networks.

4.5.3 GoogLeNet

GoogLeNet, as another representative CNN model, was inspired by network in network, especially the structure of

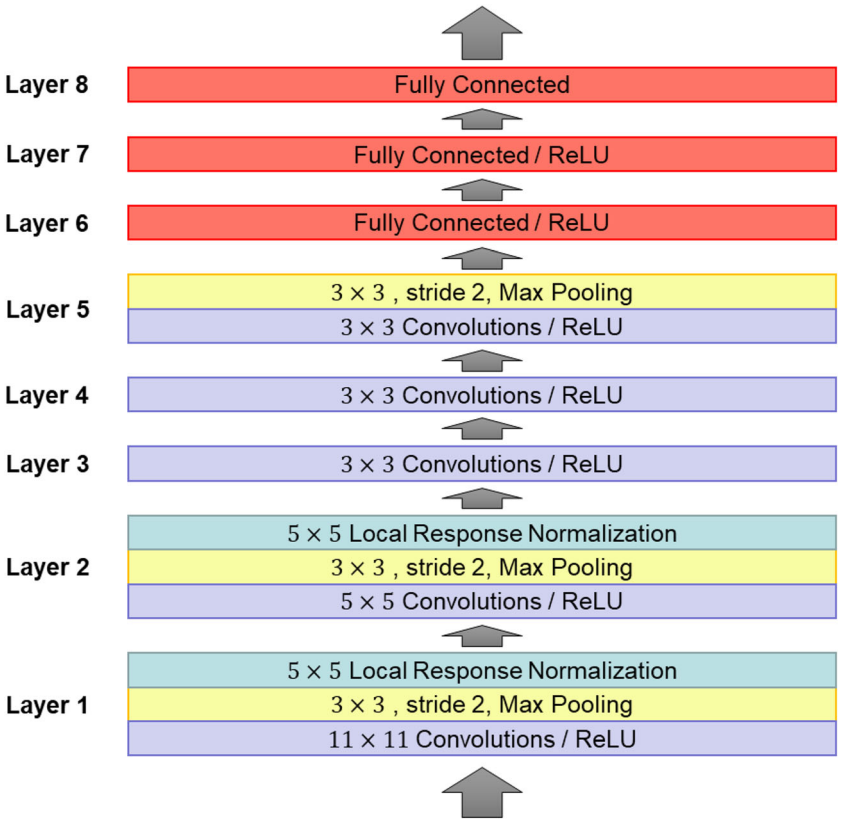

Fig. 10 Structure of AlexNet

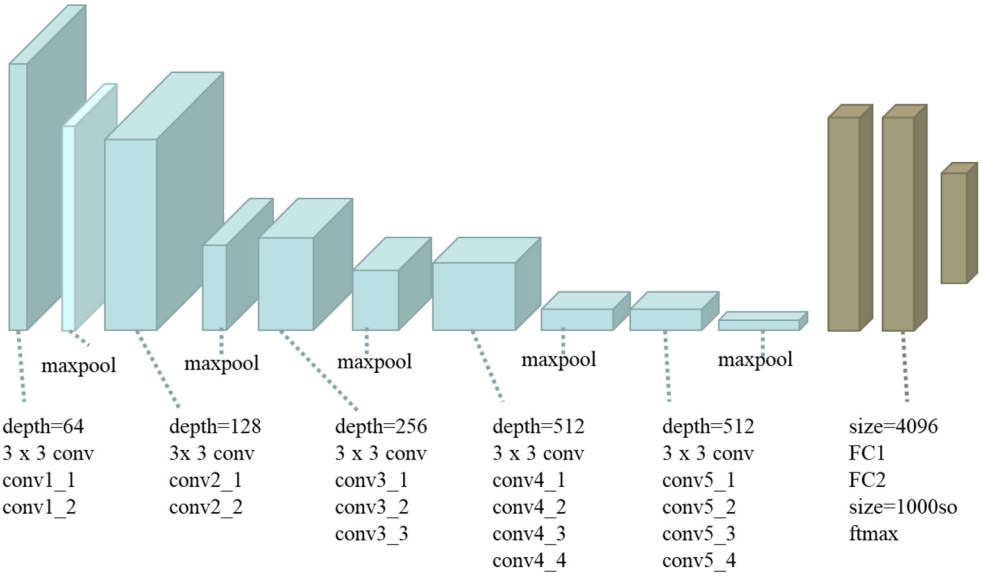

Fig. 11 Architecture of VGGNet

networks named inception. GoogLeNet is innovative in adopting the Inception structure (seeing Fig. 12) and the GAP approach. Inception architecture can broaden the width of single layer in network by using combination of 5×5, 3×3 and 1×1 three sizes of convolutional cores simultaneously in one layer to extract features and merge them into the next layer of the network in parallel with maxpooling and ReLU. The GAP approach is to overcome the problem of overfitting in fully connected layers. These ideas had a great influence on later CNN models.

4.5.4 ResNet

With the continuous development of deeper and deeper neural network's structure, CNN becomes more and more difficult to train. Worse after reaching a certain depth, if the network continues to increase its depth, the performance will not rise but fall. ResNet's first innovation was to solve the problem of deep network's structure. ResNet proposed a short-circuit-like structure (showing in Fig. 13) of connections through which multiple convolutional layers can be skipped at one time. These connections, which enable learning process to skip more than one convolutional layer at one time, can efficiently transmit gradients to very deep layers, thus breaking bottlenecks of performance in deep neural networks. ResNet's second innovation is the use of batch normalization techniques to alleviate the problem of gradient vanishing. At last, we provide Table 6 to summarize these four classic pre-trained networks and their highlights.

5 Application in Medical Image Analysis

This paper is aimed at introducing the application and development of transfer learning in the field of medical imaging. This chapter will elaborate based on medical images according to different body parts or organs.

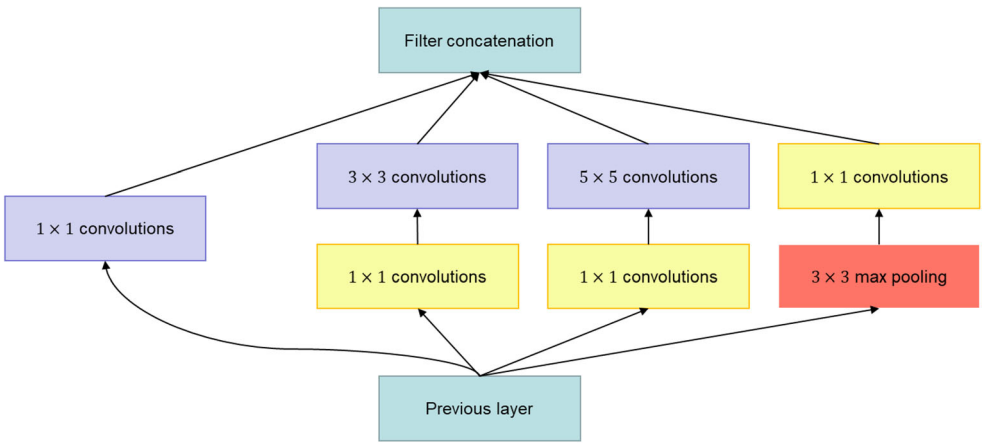

Fig. 12 Inception structure

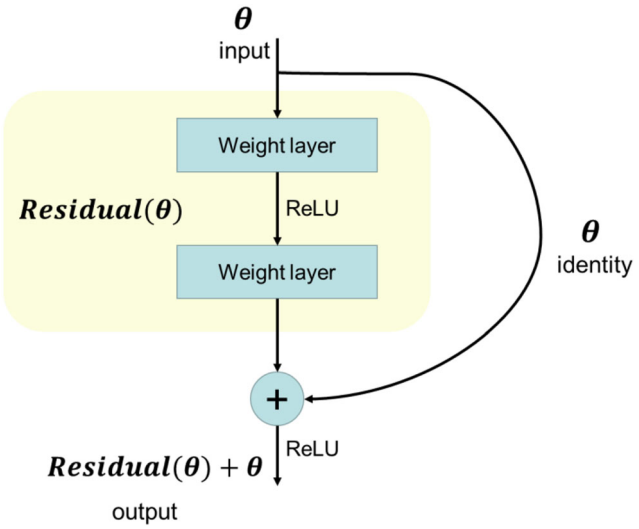

Fig. 13 Structure of residual block

5.1 Brain

At the beginning of this section, we provide Table 7 to list some of most representative papers in the field of transfer learning used in brain disease.

When transfer learning just had a great progress with the popularity of Convolutional Neural Networks, medical researchers have found its big potential in brain tumor field. Among types of brain tumors, medulloblastoma is one of most common types, about a quarter.

In [23], scientists proposed a method for medulloblastoma tumor differentiation without large amounts of labeled data using transfer learning and CNN. They noticed with transfer learning they could represent images by features learned from other different datasets. They applied VGG16 and IBCs-CNN (a CNN trained previously for invasive breast cancer tumor classification) as feature extractors and separately trained the softmax function to differentiate between anaplastic and non-anaplastic figure areas. Results told that transfer learning was superior to other methods due to its higher accuracy and lower training cost. Xu, et al. [24] took a further step proposing a terse and effective approach based on transfer learning method. They trained CNN with a huge number of images from ImageNet then transferred these extracted features (4096 neurons) in their network's architecture. And it was claimed that their approach achieved classification accuracy of 97.5% and segmentation accuracy of 84%, beating other competitive groups at that time. In addition, Ertosun and Rubin [25] also tried updating for new models from former trained networks to adapt new dataset of brain tumors. Their research showed that transfer learning and fine tuning provided capacity quickly creating new networks without beginning a new training session which was highly time-consuming.

What's more, Liu, et al. [26] succeeded in extracting features from a small dataset of brain's magnetic resonance images (MRI) using the pre-trained Convolutional Neural Network from source domain. Their approach for predicting survival time from MRI of brain tumor was based on transfer learning. Due to sizes of tumors varied, they also proposed types of image resizing methods. In this paper, they pointed that transfer learning obtained better survival time prediction with the highest accuracy of 95.4% compared to traditional methods. Besides, Saha, et al. [27] provided their modified transfer learning approach which was a model derived from previous network's construction via transferring knowledge from source domain, tackling the challenge of lacking of high-dimensional dataset when encountering survival prediction of rare cancer including brain cancer.

With the development of deep neural network and transfer learning, more and more researchers embraced transfer learning models in their schemes. Ahmed, et al. [28] pre-trained a CNN originally on ImageNet and then fine-tuned this CNN, transferring its feature learning ability to predict survival time from brain tumor MRI. They demonstrated that intentional fine-tuning could enable CNN adapt task domain where dataset was small with limited time and achieve outstanding accuracy. Lao, et al. [29] proposed a method based on transfer learning which could obtain radiomics signatures. Their research was aimed to predict overall survival of patients who suffered from Glioblastoma Multiforme (GBM), one of the most common brain tumors. They extracted 1403 handmade features and 98,304 deep features via MRI before surgery and then achieve a six-deep-feature signature, which is operated by the least absolute shrinkage and selection operator (LASSO) cox regression model. With combination of the learned signature and clinical risk factors which had been

Table 6 Summary of four classic pre-trained networks

Models	AlexNet	VGGNet	GoogLeNet	ResNet
Total number of layers(max)	8	19	22	152
Number of convolutional layers	5	16	21	151
Number of fully-connected layers	3	3	1	1
Kernel size	11 × 11, 5 × 5, 3 × 3	3 × 3	7 × 7, 1 × 1, 3 × 3, 5 × 5	7 × 7, 1 × 1, 3 × 3, 5 × 5
Top-5 error rate	16.4%	7.3%	6.7%	3.57%
Highlights	ReLU, dropout	Small kernel	Inception structure	Residual block

Table 7 Transfer learning application in brain disease

Authors	Year	Disease domain	Transfer method
Cruz-Roa, et al. [23]	2015	Medulloblastoma tumor	VGG-16 as feature extractor
Xu, et al. [24]	2015	Brain tumor	Feature extractor
Ertosun and Rubin [25]	2015	Gliomas	Fine-tuning
Liu, et al. [26]	2015	Brain tumor	Feature extractor
Saha, et al. [27]	2016	Brain tumor	Fine-tuning
Ahmed, et al. [28]	2017	Brain tumor	Fine-tuning
Lao, et al. [29]	2017	Glioblastoma multiforme	Feature extractor
Chato and Latifi [30]	2017	Brain tumor	Fine-tuning
Shen and Anderson [31]	2017	Brain tumor	Feature extractor
Ghafoorian, et al. [32]	2017	Brain lesion	Feature extractor
Li, et al. [33]	2018	Brain functional connectomes	Fine-tuning
Puranik, et al. [34]	2018	Alzheimer	Fine-tuning on Inception-V2
Rachmadi, et al. [35]	2018	Brain lesion	Fine-tuning on UNet and UResNet
Wong, et al. [36]	2018	Brain tumor	Feature extractor
Yang, et al. [37]	2018	Glioma	Fine-tuning on AlexNet and GoogLeNet
Cheng, et al. [38]	2019	Alzheimer	Feature extractor
Lu, et al. [39]	2019	Pathological brain detection	Fine-tuning on AlexNet
Talo, et al. [40]	2019	Brain abnormality classification	Fine-tuning on ResNet
Dar, et al. [41]	2020	Brain tumor	Fine-tuning
Saba, et al. [42]	2020	Brain tumor	VGG-19 as feature extractor

proved to have better performance than conventional factors, the article showed that overall survival prediction for Glioblastoma Multiforme could be firmly supported by transfer learning approach. Also focusing on prediction for overall survival of brain tumor, Chato and Latifi [30] implemented several methods including support vector machine (SVM), k-nearest neighbors (KNN), linear discriminant, tree, ensemble and logistic regression. They claimed that their classification method based on transfer learning by extracting deep features via a pre-trained CNN was the best according to experiment results. Shen and Anderson [31] also pointed that it is encouraged to deploy pre-trained CNN model on brain tumor MRI dataset when facing brain tumor MRI segmentation. If we need to adapt a current model into a new task, how much new data we should use and what portion of parameters we need to re-train are two most commonly facing problems. Trying answering these questions, Ghafoorian, et al. [32] conducted experiments by using transfer learning for MRI in brain lesion segmentation.

In 2018, Li, et al. [33] proposed their deep transfer learning neural network (DTL-NN) which could muscle classification of brain functional connection. In their work, the original task was to train a stacked sparse autoencoder (SSAE) to understand bran functional connection of healthy people and the target domain was then to transfer this model to classify some diseases or other conditions of brain functional connection like autism spectrum disorder (ASD). Compared to conventional deep neural network and SVM, their trial achieved more advanced result, including accuracy, sensitivity and specificity. Puranik, et al. [34] utilized Inception V2 model with trained knowledge of ImageNet and constructed an Alzheimer's detector based on transfer learning which presented faster and more accurate. Rachmadi, et al. [35] created a method on basis of UNet and UResNet processing output from the Irregularity Age Map (IAM) to assess brain white matter hyperintensities (WMH). It showed that transfer learning was also suitable for prediction and segmentation of brain lesion progression and regression. Wong, et al. [36] proposed a framework which used relevant data for pre-training only to learn basic shape and structure features with segmentation networks before facing real target medical classification tasks. Compared to models directly transferred to target tasks after pre-trained on ImageNet, their strategy turned to have better performance with lower computational cost on a three-class brain cancer classification problem. Yang, et al. [37] performed AlexNet and GoogLeNet with or without pre-trained on ImageNet for glioma grading and they found that transfer learning and fine-tuning could substantially promote the performance than traditional CNNs.

In the last year, the trend of transfer learning for brain medical image realm showed no sign of descending. Cheng, et al. [38] mentioned that transfer learning for early diagnosis of Alzheimer's disease usually adopted all data and annotation from source domains without discriminability of part of

irrelevant source data and labels. To tackle unreliability of source domains, their approach employed a multi-bit label coding vector instead of original binary class labels from source domains, acquiring a robust multi-label transfer feature learning (rMLTFL) model with the ability of combining features from multiple domains and kicking out those fuzzy and interfering ones. Compared with common methods, the model turned out to promote performance to significant extent. Lu, et al. [39] overcame overfitting when training neural networks to detect pathological brain on MRI images by employing AlexNet with modification of parameters, obtaining better experimental results than state-of-the-art methods. Analogously, Talo, et al. [40] trained their model based on ResNet34 with fine-tuning and optimal learning rate finder to detect brain tumors using MRI. Dar, et al. [41] demonstrated that with only tens of brain MR images, fine-tuned networks pre-trained on natural images could achieve accuracy nearly equal to networks totally trained on brain MRI datasets. It indicated that researchers could not necessarily prepare huge MRI datasets before applying deep learning into brain images, making it possible to accelerate MRI analysis. Saba, et al. [42] utilized both transfer learning model based on VGG-19 and handcrafted features by serial method and claimed to obtain quite good performance on several top brain image challenge databases. Their research showed that transfer learning with other supplementary methods [43, 44] would still have a broad prospect on brain medical image processing realm.

5.2 Heart

Before discussion of this part in detail, we first give summary of papers of transfer learning application in heart disease which is showed below in Table 8.

Since researchers began to combine medical image analysis with machine learning methods, transfer learning has always been one of hottest topics in cardiology, including cardiac diagnostics, electrocardiographic examination and cardiovascular magnetic resonance.

De Cooman, et al. [45] only used one night of patient-specific ECG data with transfer learning approach proposing a one-class support vector machine based algorithm (called TL OC-SVM) to detect epileptic seizure. They compared their method to traditional OC-SVM and concluded that with only limited specific data could transfer learning make ML classifier more accurate and robust to a certain degree. Margeta, et al. [46] proposed a CNN structure on the basis of CaffeNet which was originally trained on natural image datasets ImageNet ILSVRC2012. Their approach called CardioViewNet, aiming to recognize cardiac MRI acquisition plane, fine-tuned on CaffeNet, transferring learnt feature representatives into target domain. Compared to CNNs trained from scratch, it is claimed that transfer learning based CardioViewNet obtained better performance, promoting average F1 score to 97.66%. Al Rahhal, et al. [47] applied CNN pre-trained on ImageNet to detect Arrhythmia. After performing continuous wavelet transform (CWT) on electrocardiogram datasets, they succeeded in utilizing ECG three band images as input and making CNN completing classification. And it is indicated that transferred knowledge of reference anatomy datasets could also enhance electrocardiographic imaging prediction [48].

Murugesan, et al. [49] proposed three deep learning methods to classify cardiac arrhythmias. Among the three, their approach called ECGNet combining convolutional neural network (CNN) with long short term memory (LSTM) was the best. Moreover they proved if fixing front layers, deploying cross ECG databases and re-training on only last three layers, ECGNet could even achieve higher performance which presented promising potential of transfer learning in ECG processing. Salem, et al. [50] transformed original ECG signal data into spectrogram data which features could be extracted by pre-trained DenseNet. Alquran, et al. [51] introduced GoogLeNet and AlexNet on both bispectrum and third-order cumulants gained by input ECG data and showed that fine-tuned GoogLeNet classifier with third-order cumulants beat other state-of-the-art algorithms in precision, specificity and sensitivity when facing ECG classification. Almost toward the same goal, Byeon, et al. [52] experimented on (PTB)-ECG and their own made database, demonstrating that ResNet had better results than GoogLeNet or AlexNet when applying transfer learning strategy to ECG classification. In addition to GoogLeNet [53], AlexNet and ResNet, VGG was fine-tuned and adopted on ECG signal detection too [54]. Further, Cao, et al. [55] assembled a multi-scale advanced ResNet (MSResNet) based architecture by three fast down-sampling residual convolutional neural networks (FDResNets) independently trained of different scales. In their approach, three individual FDResNets were parallel sharing same network structure but separately trained by different scales. Then transfer learning method was implemented to inherit pre-trained weights from three source FDResNets to target MSResNet. It was given that their approach gained better performance than classic methods in atrial fibrillation detection. Due to ballistocardiogram (BCG) having less aggressiveness and being more convenient for daily monitoring than electrocardiogram (ECG), Jiang, et al. [56] utilized BCG database re-training a CNN pre-trained for ECG classification and achieved success. Van Steenkiste, et al. [57] even proved CNN structure for human ECG classification could be successfully transferred into purpose for horse ECG classification through re-training on equine electrocardiogram (eECG) database.

Apart from electrocardiogram analysis, transfer learning strategy was widely used in cardiovascular imaging field as well. Mazo, et al. [58] applied most popular convolutional neural networks including ResNet, VGG-19, VGG-16 and

Table 8 Transfer learning application in heart disease

Authors	Year	Disease domain	Transfer method
De Cooman, et al. [45]	2017	Epileptic seizure	Feature extractor
Margeta, et al. [46]	2017	Cardiopathy	Fine-tuning on CaffeNet
Al Rahhal, et al. [47]	2018	Arrhythmia	Feature extractor
Giffard-Roisin, et al. [48]	2018	Arrhythmia	Feature extractor
Murugesan, et al. [49]	2018	Arrhythmia	Fine-tuning
Salem, et al. [50]	2018	Arrhythmia	DenseNet as feature extractor
Alquran, et al. [51]	2019	Arrhythmia	Fine-tuning on GoogLeNet and AlexNet
Byeon, et al. [52]	2019	Arrhythmia	Fine-tuning on ResNet
Tadesse, et al. [53]	2019	Cardiovascular	Fine-tuning on GoogLeNet
Diker, et al. [54]	2019	Arrhythmia	Fine-tuning on VGG
Cao, et al. [55]	2019	Atrial fibrillation	Assembling Multi-scale-ResNet and Fast-downsampling-ResNets as feature extractor
Jiang, et al. [56]	2019	Atrial Fibrillation	Fine-tuning
Van Steenkiste, et al. [57]	2020	Atrial Fibrillation	Fine-tuning
Mazo, et al. [58]	2018	Cardiovascular	Fine-tuning on VGG-19, VGG-16, Inception and ResNet
Dietlmeier, et al. [59]	2019	Cardiac mitochondria	VGG-16 as feature extractor
Miyagawa, et al. [60]	2019	Vascular bifurcation	Fine-tuning

Inception as basis of transfer learning to classify cardiovascular tissues and turned to acquire remarkable results. Dietlmeier, et al. [59] adopted VGG-16 as feature extractors united with their specific designed classifier for mitochondria segmentation in cardiac cells. Miyagawa, et al. [60] changed their former work of lumen segmentation into transfer learning architecture of vascular bifurcation detection by freezing front layers of previous convolutional neural network for lumen segmentation [61, 62]. Promising capacity of transfer learning in cardiovascular imaging was pointed and researchers would go further replacing with other famous CNNs including VGG-19, GoogLeNet and ResNet according to the paper.

5.3 Breast

At first, we list some of most valued papers on transfer learning application in breast disease via Table 9.

Breast cancer is one of the severest threats for women's health. Successful therapy of breast cancer relies on early diagnosis. Computer-aided diagnosis (CADx) has a huge advantage for its efficiency and accuracy in medical imaging analysis. Transfer learning plays a significant role in CADx for breast cancer due to its benefits of no need for heavy annotation work and big database.

AlexNet was performed to extract features from digital mammographic images and showed its potential of transferring its learning ability from natural images to medical images [63]. Kandaswamy, et al. [64] applied transfer learning strategy in training procedure of convolutional neural network instead of initialization with random values. And they claimed to elevate the performance of 30% in speed and 2% in accuracy when detecting breast cancer in single-cell scale. Samala, et al. [65] utilized 2282 mammograms (source domain) to train a deep convolutional neural network as basis then froze front three layers and retrained network with 230 digital breast tomosynthesis (target domain). It is demonstrated by their experiment that transfer learning could transfer learnt related knowledge from conversant field to target field and accelerate learning process. To tackle the issue of deficient labeled data, Dhungel, et al. [66] pre-trained a deep convolutional neural network with hand-crafted features and then re-trained the network's classifier based on Inbreast database. Their fine-tuning method was presented successful in breast mass classification compared to other state-of-the-art methods in 2016. Kooi, et al. [67] also showed that training extractors on large scale related database meanwhile training classifier on limited target data could achieve comparable results compared to methods in need of considerable annotation datasets in solitary breast cysts discrimination. Samala, et al. [68] figured out a multi-task transfer learning approach for computer-aided diagnosis of breast cancer. They designed a single-task transfer learning approach as comparison as well. With single-task transfer learning approach, they just re-trained the deep convolutional neural network (DCNN) pretrained on ImageNet with only digitized screen-film mammograms (SFM) dataset. When coming for multi-task transfer learning approach, they re-trained the DCNN with SFM dataset, Digital Database for Screening Mammography (DDSM) dataset and digital mammograms (DM) dataset. Through multiple learning task including assistant tasks' learning, it is illustrated that multi-task transfer learning method presented

Table 9 Transfer learning application in breast disease

Authors	Year	Disease domain	Transfer method
Huynh, et al. [63]	2016	Mammographic Tumor	AlexNet as Feature extractor
Kandaswamy, et al. [64]	2016	Breast Cancer	Fine-tuning
Samala, et al. [65]	2016	Breast Tomosynthesis	Fine-tuning
Dhungel, et al. [66]	2017	Masses in mammograms	Fine-tuning
Kooi, et al. [67]	2017	Mammography	Feature extractor
Samala, et al. [68]	2017	Mammograms	Fine-tuning (Multi-task)
Yap, et al. [69]	2017	Breast Lesions	Fine-tuning
Chougrad, et al. [70]	2018	Breast cancer	Fine-tuning
Mohamed, et al. [71]	2018	Mammographic	Fine-tuning on AlexNet
Samala, et al. [72]	2018	Digital breast tomosynthesis	Fine-tuning
Samala, et al. [73]	2018	Digital breast tomosynthesis	Feature extractor
Zhang, et al. [74]	2018	Breast cancer	Feature extractor
Byra, et al. [75]	2019	Breast mass classification	Feature extractor & Fine-tuning
Khan, et al. [76]	2019	Breast cancer	Feature extractor
Mendel, et al. [77]	2019	Digital breast tomosynthesis	Feature extractor
Xie, et al. [78]	2019	Breast histopathological images	Inception_ResNet_V2 as Feature extractor
Yu, et al. [79]	2019	Mammographic breast lesions	Feature extractor
Zhu, et al. [80]	2019	Radiogenomic associations in breast cancer	Feature extractor
Zhu, et al. [81]	2019	Breast MRIs	Feature extractor

better performance in generalization and accuracy than single-task transfer learning method. Yap, et al. [69] implemented three independent deep convolutional neural networks to detect breast lesions, containing a patch-based LeNet, a UNet and an FCN-AlexNet which was pre-trained with transfer learning technology. It was indicated that transfer learning based approach obtained best scores on more than half of datasets.

In 2018, Chougrad, et al. [70] proposed a deep convolutional neural network to classify mammography mass lesion. Further, they discussed several issues such as depth and architecture of network, data, whether to perform transfer learning and how these issues could affect network's performance. It is demonstrated that first rather than initializing the deep convolutional network with random values, initialization with pre-trained network's values seems a better choice. Second, it is not always getting better performance when doing more fine-tuning jobs. Overfitting would increasing likely occur if overusing fine-tuning especially for deep network structure and insufficient data. Mohamed, et al. [71] designed a deep convolutional neural network for classification of breast mammogram density categories. Beside the proposed CNN architecture, they also deployed a transfer learning approach using a modified AlexNet pre-trained on ImageNet then fine-tuning it on breast mammographic images database as comparison. Via their study, it was pointed that transfer learning could achieve almost equivalent performance to delicately designed CNN structure but with largely reduction in training cost. And another benefit of transfer learning is its accuracy seems not to depend on the quantity of training samples in fine-tuning procedure, which means pre-trained CNN has already sufficient capacity to extract and represent features of breast mammograms even it was originally trained on natural images. Similarly, Samala, et al. [72] experimented on transfer learning strategies for breast cancer detection. The first transfer learning strategy was a one-stage transfer learning method, fixing first convolutional layer, first pooling layer and first normalization layer then re-training other layers together using mammograms (SFM&DM) based on pre-trained AlexNet. The second transfer learning method, which was also called stage-two, was to freeze all layers except for last fully connected layer of stage-one networks and then re-train the networks on Digital breast tomosynthesis (DBT) datasets. In this paper, they aimed to discover impact to transfer learning strategies on variant scale of training samples. It was concluded that applying multi-stage transfer learning could achieve prominent promotion, being a more efficient way compared to simply increasing the quantity of training samples. Further, in another article [73], they performed nearly same deep convolutional neural network structure and revealed that using multi-stage transfer learning structure as feature extractors and deploying pathway evolution of feature selection and random forest classification could significantly decrease the numbers of neurons and parameters in networks, making transfer learned deep convolutional neural network more neat and effective [82, 83]. Moreover, transfer learning

strategy was tested and verified in photoacoustic images of breast cancer as well [74].

In 2019, Byra, et al. [75] went deeper in exploiting transfer learning strategies for breast mass classification with sonography images. They performed three different transfer learning methods and compared each performance on two public datasets. Beginning with the first approach, they applied VGG-19 as feature extractors with aiding the classifier of support vector machine (SVM). As for the second method, they fine-tuned the VGG-19 with fixing the first four convolutional blocks only fine-tuning on the fifth convolutional block and fully connected layers to acquire optimal performance. Last, they proposed a novel transfer learning based deep convolutional neural network method by introducing matching layer. Matching layer was used between input raw images and pre-trained VGG-19 convolutional blocks, converting grayscale images to RGB ones making images augmented to tap potential of DCNN feature extractors. With the same purpose of maximizing the use of deep convolutional feature extractors, Khan, et al. [76] introduced data augmentation technology to transfer learning method in breast cancer detection. Data augmentation is an approach to perform image processing such as rotation, coloring, scaling and transformation, in which datasets of target images could be enlarged. With enriched samples of data, it could make feature extractor more powerful and mitigate over-fitting [84, 85]. Another novelty of this paper is to apply average pooling instead of fully connected layers as classifier.

More researchers deployed kinds of experiments looking for further possibilities of transfer learning technique on breast images analysis in different aspects. Mendel, et al. [77] studied the effect of different ways of breast cancer screening on the performance when applying deep learning method with transfer learning strategy. Based on their work, it is concluded that digital breast tomosynthesis (DBT) excelled at enabling pre-trained convolutional neural networks to maximize its strength as feature extractors compared to traditional full-field digital mammography (FFDM). Xie, et al. [78] not only demonstrated that based on Inception_ResNet_V2, transfer learning method could achieve best performance in breast histopathological images analysis within supervised learning field, but also showed its superiority in unsupervised learning of extracting features for their proposed new autoencoder transforming these features to lower dimensional space in purpose of clustering. And of course, scientists still tried exploiting potentials of state-of-the-art convolutional neural network architectures [79–81] as pre-trained models applied in transfer learning to tackle the issues of breast images analysis.

5.4 Lung

Before discussion, we give summary of selected papers on transfer learning application in lung disease through Table 10.

Table 10 Transfer learning application in lung disease

Authors	Year	Disease domain	Transfer method
Sawada and Kozuka [86]	2015	Lung CT	Fine-tuning
Shouno, et al. [87]	2015	Diffuse lung disease	Fine-tuning
Christodoulidis, et al. [88]	2016	Lung CT	Feature extractor
Paul, et al. [89]	2016	Survival prediction of lung adenocarcinoma	VGG as Feature extractor
Seelan, et al. [90]	2016	Lung Lesion	Feature extractor
Shen, et al. [91]	2016	lung cancer prediction	Feature extractor
Nibali, et al. [92]	2017	Pulmonary nodule classification	Fine-tuning on ResNet
Hussein, et al. [93]	2017	Lung nodule classification	Feature extractor
Shan, et al. [94]	2017	Lung nodule classification	Feature extractor
Wang, et al. [95]	2017	Lung nodule classification	Feature extractor
da Nóbrega, et al. [96]	2018	Lung nodule classification	Feature extractor
Hosny, et al. [97]	2018	Lung cancer prognostication	Feature extractor
Dey, et al. [98]	2018	Lung nodule classification	Fine-tuning on DenseNet
Fang [99]	2018	Lung nodule classification	Fine-tuning on GoogLeNet
Nishio, et al. [100]	2018	Lung nodule classification	VGG-16 as feature extractor
Hussein, et al. [101]	2019	Lung cancer	Feature extractor
Lakshmi, et al. [102]	2019	Lung carcinoma	VGG-16 & VGG-19 as feature extractor
Li, et al. [103]	2019	Lung nodule detection	Fine-tuning
Shi, et al. [104]	2019	Lung nodule detection	Fine-tuning on VGG-16
Zhang, et al. [105]	2019	Lung nodule detection	Fine-tuning on LeNet-5
Huang, et al. [106]	2020	Lung nodule detection	Feature extractor followed by extreme learning machine

It has a long history since transfer learning was introduced to analyzing medical images of lung for aiding doctors' diagnosis and treatment. Sawada and Kozuka [86] trained multi-prediction deep Boltzmann machine (MPDBM) on natural image database to satisfy requirements in source domain. Then they transferred this pre-trained networks and fine-tuned it using X-ray CT images of lung in target domain to solve classification. Shouno, et al. [87] demonstrated that it is a more efficient way to apply deep convolutional neural network as pre-trained network with training on non-medical images for diffuse lung diseases (DLD) on high-resolution computed tomography (HRCT) images. Compared to training from scratch on DLD-HRCT images or on natural images only, transfer learning strategy, which is to pre-train on natural images first then transfer learnt knowledge to DLD-HRCT domain, could achieve better performance. Christodoulidis, et al. [88] adopted six public texture images databases to enhance CNN's ability of extracting low-level features. After pre-trained on texture databases, the convolutional neural network architecture turned to increase accuracy by 2% in classification of lung CT scanning images. Paul, et al. [89] utilized pre-trained convolutional neural networks to extract deep features combining with conventional hand-crafted features. They applied these features with kinds of classifiers and claimed that the assembly of pre-trained VGG-f structure and symmetric uncertainty feature ranking algorithm followed by a random forests classifier could obtain best results in predicting survival time of lung cancer patients. It is concluded that transfer learning could help doctors and researchers extract deep features of lung scanning images from learnt knowledge of source domain [90, 91].

Nibali, et al. [92] fine-tuned a pre-trained ResNet to adapt pulmonary classification. Hussein, et al. [93] proposed a 3D based convolutional neural networks architecture for lung nodules recognition. In order to improve the ability of acquiring deep features more representatively, they pre-train the proposed 3D CNN architecture with non-medical images before applying it to lung nodules CT images. Using transfer learning strategy that employing pre-trained convolutional neural networks as feature extractors in lung radiography has been a trend [94–97]. Dey, et al. [98] and Fang [99] separately fine-tuned 3D DenseNet and GoogLeNet with chest three dimensional CT scans for lung nodules classification and both of them claimed to achieved state-of-the-art results based on transfer learning strategy. Nishio, et al. [100] adopted VGG-16 as feature extractor in lung nodules classification and they demonstrated that transfer learning method was superior to hand-crafted features and traditional machine learning method like SVM. And it is showed that input images of bigger scale could elevate deep convolutional neural network structure's performance. Hussein, et al. [101] applied deep convolutional neural networks with transfer learning method in risk stratification of lung cancer. Apart from approval of pre-trained DCNN in supervised learning, they also pointed its deficiency compared to proportion-support vector machine in unsupervised learning for transfer learning still cannot get rid of annotated training data. Lakshmi, et al. [102] demonstrated that VGG-16 and VGG-19 could be applied with transfer learning method to detect the lung carcinomas tissues in deficiency of annotated images.

In spite of large efforts have been devoted to implementing deep learning to detect lung nodule from CT images, it still exists potential for deploying CNN based method on lung MR images. Li, et al. [103] proposed a method using transfer learning strategy to fine-tune a faster R-CNN targeting region of interest of lung nodule and showed it could obtain good accuracy compared to traditional machine learning method with quick neural networks construction progress. As we know that acquiring low false positive rate is challenging to most of lung nodules detection method when processing thoracic computed tomography. Shi, et al. [104] contributed a deep learning based method in their latest paper to reduce FP rate of nodule detection. They fine-tuned VGG-16 with transfer learning and modified parameters in fully connected layers to make networks more efficient for nodule detection. Then they utilized this fine-tuned networks as feature extractors to extract features of lung nodules and trained a support vector machine (SVM) for nodule classification. Zhang, et al. [105] performed their experiment along with this research direction but using LeNet-5 as basic convolutional neural networks for fine-tuning rather than VGG structure. Further, Huang, et al. [106] proposed a new method combining deep transfer convolutional neural networks with extreme learning machine to diagnose lung nodules based on CT images. They applied deep transfer CNN extracting high level features of lung nodules and then followed by an extreme learning machine classifier. It was showed to obtain result better than other state-of-the-art methods.

5.5 Kidney

At the head of this part, we provide Table 11 to summarize important papers on transfer learning application in kidney disease.

As one of the most important organs in the human body, kidney is very vulnerable to various diseases. Medical research on kidney has been a hot topic since many years ago. Back to 2013 when transfer learning began to be introduced to public, researchers started using transfer learning technology aiding kidney diagnosis [107]. Wankhade and Patey proposed a bisecting k-means algorithm based on patients' test report aiming to predict several diseases including kidney. They showed that knowledge of kidney research and diagnosis was transferable and transfer learning could be applied in kidney diagnosis with promising future. And it has been proved correct. Since deep learning method has been most popular

Table 11 Transfer learning application in kidney disease

Authors	Year	Disease domain	Transfer method
Wankhade and Patey [107]	2013	Glomeruli classification	Feature extractor
Marsh, et al. [108]	2018	Glomeruli classification	Fine-tuning
Zheng, et al. [109]	2018	Glomeruli classification	AlexNet as Feature extractor
Zheng, et al. [110]	2019	Glomeruli classification	Comparison between Feature extractor and fine-tuning
Efremova, et al. [111]	2019	Kidney segmentation	Feature extractor
Hao, et al. [112]	2019	Chronic kidney disease	Feature extractor
Kannan, et al. [113]	2019	Glomeruli segmentation	Fine-tuning
Kuo, et al. [114]	2019	Kidney function prediction	ResNet as feature extractor
Wu, et al. [115]	2019	Kidney ultrasound pathology	Feature extractor
Yin, et al. [116]	2018	Kidney segmentation	Feature extractor
Yin, et al. [117]	2020	Renal ultrasound images analysis	Feature extractor
Ayyar, et al. [118]	2019	Glomeruli classification	Inception_ResNet_V2 as feature extractor
Mathur, et al. [119]	2020	Glomeruli classification	Multi-Gaze Attention Networks as feature extractor

methods of transfer learning technology, more and more researchers utilized transfer learning for the purpose of kidney diagnosis and disease prediction. Marsh, et al. [108] applied transfer learning strategy in classification of non-sclerosed and sclerosed glomeruli of frozen section biopsies to help doctors estimate whether the kidney is transplantable. Their research has concluded that pre-trained Convolutional Neural Network could obtain better performance than method directly trained on small dataset of glomeruli. Zheng and his colleagues proposed a transfer learning based method to classify ultrasound images of kidneys in order to build aiding diagnosis system for congenital abnormalities of the kidney and urinary tract (CAKUT) in children [109, 110]. They compared different strategies of transfer learning based on AlexNet and have demonstrated that features extracted by AlexNet combined with features of hand-crafted augmented with Support Vector Machine (SVM) as classifier obtained most ideal performance on their dataset, compared to both fine-tuning pure transfer learning method and conventional SVM method. Efremova, et al. [111] also experimented transfer learning method on automatic segmentation of kidney and liver CT images and achieved outstanding competition scores in the 2019 Kidney Tumor Segmentation (KiTS-2019) challenge.

Hao, et al. [112] proposed a novel transfer learning method for screening of chronic kidney disease based on ultrasound images. They applied the pre-trained deep CNN structure extracting both texture features and deep features then using these mixed features for classification and achieved convincing result on dataset consisting of 226 ultrasound images. Kannan, et al. [113] embraced transfer learning method by deploying a pre-trained convolutional neural network as basis then retraining their networks on three categories of trichrome images to perform glomeruli segmentation. Kuo, et al. [114] proposed their transfer learning-based method blended with several deep learning schemes for intelligently predicting the estimated glomerular filtration rate (eGFR) and chronic kidney disease (CKD) status. They first applied ResNet pre-trained on ImageNet as feature extractors. And they utilized data augmentation to enrich annotated information, applied kidney length annotations to mark key region of kidney and employed bootstrap aggregation to alleviate overfitting and enhance capacity of generalization. It was demonstrated that their deep transfer learning method had beaten human eyes by obtaining accuracy 85.6% higher than experienced doctors. Wu, et al. [115] proposed their novel merged deep framework structure called PASnet for analysis of kidney ultrasound images. They combined pre-trained deep convolutional neural network with Siamese network joint training and showed this mixed structure could adopt advantages of both two networks and achieved better performance than either of them. Yin, Peng and their team [116, 117] contributed to renal ultrasound images segmentation by introducing subsequent boundary distance regression and pixel classification. They utilized pre-trained deep convolutional neural network extracting features then applied these features as input to calculate kidney boundary distance maps through boundary distance regression network. Aided with pixel classification network to distinguish renal pixels from non-renal pixel and data augmented scheme, they demonstrated that their method could obtain encouraging performance for automatic kidney segmentation.

Other researchers also found transfer learning based method did not always result in reliable success in any condition. Ayyar, et al. [118] adopted a novel medical image dataset which had binary classes (normal and abnormal) of renal glomeruli images called Glomeruli Classification Database (GCDB). They performed popular transfer learning algorithms on the dataset to do binary classification of glomeruli. It was demonstrated that not any transfer learning based

method could achieve satisfied result. For example, ResNet50 and InceptionV3 even could not obtained good performance compared to traditional pre-trained image classifier. The best performance on this dataset was acquired by logistic regression model enhanced with feature extractors from Inception_ResNet_V2. Further [119], they keep exploring glomeruli classification by performing experiments on one of the latest deep learning architectures, Multi-Gaze Attention Networks, and claimed to obtain state-of-the-art performance.

6 Discussion

It is obvious that transfer learning has achieved great progress in medical image analysis within many fields and tasks. However, some problems or difficulties would still be met when facing real medical image analysis tasks. Among these problems, some may have been overcome or mitigated with solution while others have not. So, in next paragraphs we will discuss on this topic.

6.1 Data

As we all know, medical images are always hard and expensive to collect compared to other ordinary vision tasks. And some kinds of medical image capturing methods even do harm to patients' bodies, which means we cannot obtain large amount of data naturally [120]. Thus, in most cases, when we perform transfer learning in medical image analysis, we are only equipped with very limited data in target domain. To tackle with this problem, we mainly focus on two direction of solutions. First, although the quantity of medical image dataset is usually small, we may have method to enrich it, which leads us to the technology called data augmentation. Second, even with limited data, we need to make the best of it, avoiding condition of pool quality of images. Another technique called smart imaging is aimed at enhancing quality of required images, which is just what we need. Therefore, in this section, we will give a brief discussion about data augmentation and smart imaging.

6.1.1 Data Augmentation

The comforting thing is that we have an easy but effective approach called data augmentation. Basically, data augmentation is to perform photometric and geometric transformations on original images such as scaling, rotation and reflection. In this way, data augmentation dilates the amount of original image datasets and enriches its distribution making it more close to the distribution in the real world. Data augmentation has been a common approach before adopting transfer learning and could been seen applied in more and more relevant works [121]. It is admitted being one of the most effective and practical ways for relieving the scarcity of data.

6.1.2 Smart Imaging

Besides, smart imaging is another direction we should pay attention to. Smart imaging is applied for obtaining image data with better quality. It could help increase images' resolution, detect shadow, and reduce noise and artifacts [122]. All these impacts enable deep learning algorithms to result in higher accuracy and faster procedure.

6.2 Labels

Handling medical image analysis, lacking data maybe not the most intractable problem while lacking labels maybe. To get rid of lacking labeled data, it is natural to consider bifurcate two ways. The first way is admitting labeled data is limited; thus, we shall find method to produce more labeled data, better if the producing progress is automatic. The way of this is called self-supervised learning. The second method we need consider is that given limited labeled data, how to perform transfer learning just with it. This method is equal to finding a strategy which could relieve the dependency of labeled data during transfer learning as much as possible. And that refers to technique called unsupervised domain adaption. We will discuss self-supervised learning and unsupervised domain adaption separately in following parts.

6.2.1 Self-Supervised Learning

Recently, a novel strategy named self-supervised learning came out. Self-supervised learning could produce labels with non-labeled original data from scratch without human annotation by designing and performing some artificial tasks with no actual use. For example, predicting rotation angle and relative position of patches in image are two of the most common and intentional tasks designed in self-supervised learning [123]. Through these contrived learning process, automatically generated labels without human intervention could be acquired. We believe it will be a trend and seen in more and more papers on medical images processing in future.

6.2.2 Unsupervised Domain Adaption

Self-supervised learning is suitable for condition with limited or none labels in target domain while performing transfer learning. But if labels in source domain are sufficient, we could apply another method called unsupervised domain adaption. Unlike self-supervised learning, unsupervised domain adaption does not produce labels but eliminate the dependence on labels in target domain. The most popular unsupervised domain adaption now is adversarial-based [124]. In

general, the purpose of adversarial-based unsupervised domain adaption is to map data from source domain and target domain into a same feature space, making the distance between two domains in feature space as close as possible. Therefore, the task function originally trained on source domain will be easily transferred to target domain and achieves ideal result. More specifically, adversarial-based unsupervised domain adaption consists of three components, a feature extractor G_f, a label predictor G_y and a domain classifier G_d. These three fuction each has its different role. G_f extracts features from input images, mapping data to feature space so that G_y can perform classification correctly based on labels in source domain while G_d cannot differentiate if the feature data comes from source domain or target domain. G_y aims at outputting labels of source domain as accurately as possible. G_d needs to tell whether the data of the feature space is from source domain or target domain. In the end, we hope to reach the condition that G_d cannot differentiate feature source which means data from source domain and data from target domain nearly share same distribution in feature space. Thus, G_y could be applied as effective classifier within target domain and the problem of lacking labels in target domain is also overcome.

6.3 Models

Talking about transfer learning, it does not consist of data and labels only. Models play the same important role in transfer learning as data's and label's. During its development history, transfer learning has turned to disclose a trend of interacting with other cutting-edge concepts in deep learning and being applied with combination of other deep learning based models more and more often. Under this background, we will discuss this trend in three aspects. First is how the theory of few shot learning and meta learning impacts on transfer learning. Second, we discuss transfer learning's combination with other popular deep learning models. And third, it is usual but indispensable that we need to talk about model interpretability of transfer learning, which maybe an existing pale aspect of transfer learning.

6.3.1 Few Shot Learning and Meta Learning

As we have mentioned many times, medical image analysis is always accompanied by a shortage of samples and labels. In other words, most of medical image analysis tasks belong to few shot learning realm. Meta learning has been recognized as most successful and promising method to solve few shot learning problems. Note that meta learning shares some similarities in thoughts and procedure with transfer learning. In typical transfer learning, researchers fine-tuned a pre-trained convolutional neural network or regard it as feature extractor followed by other classifiers. Meta learning could be divided into three categories, learning to fine-tuning, RNN memory based and metric learning. The first category also the most common category of meta learning, learning to fine-tuning, could be compared to fine-tuning in transfer learning in many aspects. Learning to fine-tuning refers to learning a good initialization parameter and when facing new tasks, with only few shot samples the learnt initialization parameter could achieve ideal performance after few steps of gradient descent procedure. Another method of learning to fine-tuning is to train an optimizer based on like LSTM for assisting fine-tuning. Therefore, it is convinced that fine-tuning is widely applied in meta learning. Although articles of medical image analysis on few shot learning or meta learning rarely come out for now, we believe it is a trend and maybe in future the concept of transfer learning would merge in the concept of meta learning.

6.3.2 Combination with Other Deep Learning Models

Since scientists utilized pre-trained convolutional neural network as feature extractor, transfer learning has been combined with other deep learning approaches for a long time. For example, as we addressed above, adversarial-based unsupervised domain adaption applied adversarial network and training process which originated form generative adversarial network (GAN). It is a representative example of transfer learning's combination with other deep learning models. In fact, some researchers have tried adopting transfer learning strategy in reinforcement learning [125]. It is the same that though few papers of medical image on this field have been published, we believe they will be in near future.

6.3.3 Model Interpretability

Last but not the least, the problem that lack of impeccable interpretation of transfer learning and its variety of pre-trained models, still remains. Although some papers have explored in this field but few of them were addressed from the view of medical image analysis. It has become a cliché, but we think it deserves mentioning again. And we are looking forward to seeing more works and articles focusing on this realm.

7 Conclusion

In this survey article, we first introduce the main research issues in the field of medical image analysis and the history of transfer learning in it. Then we review the elemental theory and development of convolutional neural network and transfer learning. After that we select brain, heart, breast, lung and kidney, five common fields of medical imaging research, and list the representative papers in detail and summarize their methods. Finally, we discuss the future development trend of transfer learning in the field of medical imaging and some

possible combination directions [126, 127]. We believe that in the field of medical image research, transfer learning will gradually develop towards meta learning in combination with technologies such as data augmentation, self-supervised learning and domain adaptation. Or transfer learning will combine with reinforcement learning and other models to produce some more effective and powerful models to comprehensively improve the performance of neural networks at the current level. The author's knowledge is limited, perhaps some important works are not included in this article. However, we hope that this survey paper will provide a positive and illuminating perspective to review the development and trends of transfer learning in the field of medical imaging.

Acknowledgements This study is partially supported by Royal Society International Exchanges Cost Share Award, UK (RP202G0230); Medical Research Council Confidence in Concept Award, UK (MC_PC_17171); Hope Foundation for Cancer Research, UK (RM60G0680); Fundamental Research Funds for the Central Universities (CDLS-2020-03); Key Laboratory of Child Development and Learning Science (Southeast University), Ministry of Education.

References

1. Chen H, Qi X, Yu L, Dou Q, Qin J, Heng P-A (2017) DCAN: Deep contour-aware networks for object instance segmentation from histology images. Med Imag Anal 36:135–146
2. Caravagna G, Giarratano Y, Ramazzotti D, Tomlinson I, Graham TA, Sanguinetti G, et al. (2018) Detecting repeated cancer evolution from multiregion tumor sequencing data. Nature Methods vol. 15, pp. 707–714
3. Du Y et al (2018) Classification of tumor epithelium and stroma by exploiting image features learned by deep convolutional neural networks, 46, 12, 1988–1999
4. Shin H-C et al (2016) Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning, 35, 5, 1285–1298
5. Armato SG et al (2011) The lung image database consortium, (LIDC) and image database resource initiative (IDRI): a completed reference database of lung nodules on CT scans. Med Phys 38(2):915–931
6. Aerts H et al (2014) Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. Nat Commun 5 Art. no. 4006
7. Chollet F (2017) and Ieee, Xception: Deep Learning with Depthwise Separable Convolutions. In: 30th Ieee conference on computer vision and pattern recognition (IEEE Conference on Computer Vision and Pattern Recognition, pp 1800–1807
8. Cover TM, Hart PE (1967) Nearest neighbor pattern classification. IEEE Trans Inf Theory 13(1):21
9. Cheng PM, Malhi HS (2017) Transfer learning with convolutional neural networks for classification of abdominal ultrasound images. J Digit Imaging 30(2):234–243
10. Tajbakhsh N, Shin JY, Gurudu SR, Hurst RT, Kendall CB, Gotway MB, Liang J (2016) Convolutional neural networks for medical image analysis: full training or fine tuning? IEEE Trans Med Imaging 35(5):1299–1312
11. Chang H, Han J, Zhong C, Snijders AM, Mao J-H, M. intelligence (2017) Unsupervised transfer learning via multi-scale convolutional sparse coding for biomedical applications. IEEE Trans Pattern Anal Mach Intell 40(5):1182–1194
12. Liu S, Liu G, Zhou H (2019) A robust parallel object tracking method for illumination variations. Mob Netw Appl 24(1):5–17
13. Liu S, Liu X, Wang S, Muhammad K (2020) Fuzzy-aided solution for out-of-view challenge in visual tracking under IoT-assisted complex environment. Neural Comput Applic
14. Liu S, Guo C, Al-Turjman F, Muhammad K, de Albuquerque VHC (2020) Reliability of response region: A novel mechanism in visual tracking by edge computing for IIoT environments. Mech Syst Signal Process 138:106537
15. Liu S, Lu MY, Li HS, Zuo YC (2019) Prediction of gene expression patterns with generalized linear regression model (in English). Front Genet 10:11 Art. no. 120
16. Liu S, Chen X, Li Y, Cheng XC (2019) Micro-distortion detection of lidar scanning signals based on geometric analysis (in English). Symmetry-Basel 11(12):13 Art. no. 1471
17. Huang C et al A dynamic priority strategy for IoV data scheduling towards key data
18. Chenxi H et al (2020) Sample imbalance disease classification model based on association rule feature selection
19. Saxe AM, McClelland JL, Ganguli S (2013) Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. arXiv preprint arXiv:1312.6120
20. Glorot X, Bengio Y (2010) Understanding the difficulty of training deep feedforward neural networks. In: Proceedings of the thirteenth international conference on artificial intelligence and statistics, pp 249–256
21. He K, Zhang X, Ren S, Sun J (2015) Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In: Proceedings of the IEEE international conference on computer vision, pp 1026–1034
22. LeCun YA, Bottou L, Orr GB, Müller K-R (2012) Efficient backprop. In: Neural networks: Tricks of the trade: Springer, pp 9–48
23. Cruz-Roa A, Arévalo J, Judkins A, Madabhushi A, González F (2015) A method for medulloblastoma tumor differentiation based on convolutional neural networks and transfer learning. In: 11th International Symposium on Medical Information Processing and Analysis, vol 9681, p 968103: International Society for Optics and Photonics
24. Xu Y et al (2015) Deep convolutional activation features for large scale brain tumor histopathology image classification and segmentation. In: 2015 IEEE international conference on acoustics, speech and signal processing (ICASSP), pp 947–951: IEEE
25. Ertosun MG, Rubin DL (2015) Automated grading of gliomas using deep learning in digital pathology images: a modular approach with ensemble of convolutional neural networks. In: AMIA Annual Symposium Proceedings, vol 2015, p 1899: American Medical Informatics Association
26. Liu R, Hall LO, Goldgof DB, Zhou M, Gatenby RA, Ahmed KB (2016) Exploring deep features from brain tumor magnetic resonance images via transfer learning. In: 2016 International Joint Conference on Neural Networks (IJCNN), pp 235–242: IEEE
27. Saha B, Gupta S, Phung D, Venkatesh S (2016) Transfer learning for rare cancer problems via discriminative sparse gaussian graphical model. In: 2016 23rd International Conference on Pattern Recognition (ICPR), pp 537–542: IEEE
28. Ahmed KB, Hall LO, Goldgof DB, Liu R, Gatenby RA (2017) Fine-tuning convolutional deep features for MRI based brain tumor classification. In: Medical Imaging 2017: Computer-Aided Diagnosis, vol 10134, p 101342E: International Society for Optics and Photonics
29. Lao J et al (2017) A deep learning-based radiomics model for prediction of survival in glioblastoma multiforme, 7, 1, 10353

30. Chato L, Latifi S (2017) Machine learning and deep learning techniques to predict overall survival of brain tumor patients using MRI images. In: 2017 IEEE 17th International Conference on Bioinformatics and Bioengineering (BIBE), pp 9–14: IEEE
31. Shen L, Anderson T (2017) Multimodal brain MRI tumor segmentation via convolutional neural networks, ed
32. Ghafoorian M et al (2017) Transfer learning for domain adaptation in mri: Application in brain lesion segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp 516–524
33. Li H, Parikh NA, He L (2018) A novel transfer learning approach to enhance deep neural network classification of brain functional connectomes. Front Neurosci 12
34. Puranik M, Shah H, Shah K, Bagul S (2018) Intelligent Alzheimer's Detector using Deep Learning and IEEE (Proceedings of the 2018 Second International Conference on Intelligent Computing and Control Systems). IEEE, New York, pp 318–323
35. Rachmadi MF, Valdés-Hernández MdC, Komura T (2018) Transfer Learning for Task Adaptation of Brain Lesion Assessment and Prediction of Brain Abnormalities Progression/Regression using Irregularity Age Map in Brain MRI. In: International Workshop on PRedictive Intelligence In MEdicine, Springer, pp 85–93
36. Wong KCL, Syeda-Mahmood T, Moradi M (2018) Building medical image classifiers with very limited data using segmentation networks (in English). Med Image Anal 49:105–116
37. Yang Y et al (2018) Glioma grading on conventional MR images: a deep learning study with transfer learning. Front Neurosci 12
38. Cheng B, Liu M, Zhang D, Shen D (2019) Robust multi-label transfer feature learning for early diagnosis of Alzheimer's disease. Brain Imaging Behav 13(1):138–153
39. Lu S, Lu Z, Zhang Y-D (2019) Pathological brain detection based on AlexNet and transfer learning. J Comput Sci 30:41–47
40. Talo M, Baloglu UB, Yıldırım Ö, Rajendra Acharya U (2019) Application of deep transfer learning for automated brain abnormality classification using MR images. Cogn Syst Res 54:176–188
41. Dar SUH, Özbey M, Çatlı AB, Çukur T (2020) A transfer-learning approach for accelerated MRI using deep neural networks. Magn Reson Med 84:663–685
42. Saba T, Sameh Mohamed A, El-Affendi M, Amin J, Sharif M (2020) Brain tumor detection using fusion of hand crafted and deep learning features. Cogn Syst Res 59:221–230
43. Li JP, Qiu S, Shen YY, Liu CL, He HG (2020) Multisource transfer learning for cross-subject EEG emotion recognition. IEEE Trans Cybern 50(7):3281–3293
44. Tandel GS, Balestrieri A, Jujaray T, Khanna NN, Saba L, Suri JS (2020) Multiclass magnetic resonance imaging brain tumor classification using artificial intelligence paradigm. Comput Biol Med 122 Art. no. 103804
45. De Cooman T, Varon C, Van de Vel A, Ceulemans B, Lagae L, Van Huffel S (2017) Semi-supervised one-class transfer learning for heart rate based epileptic seizure detection, in 2017 Computing in Cardiology (CinC), pp. 1–4: IEEE
46. Margeta J, Criminisi A, Cabrera Lozoya R, Lee DC, Ayache N (2017) Fine-tuned convolutional neural nets for cardiac MRI acquisition plane recognition. Comput Methods Biomech Biomed Eng: Imaging Vis 5(5):339–349
47. Al Rahhal MM, Bazi Y, Al Zuair M, Othman E, BenJdira BJJOM (2018) Convolutional neural networks for electrocardiogram classification. B. Eng 38(6):1014–1025
48. Giffard-Roisin S et al (2018) Transfer learning from simulations on a reference anatomy for ECGI in personalized cardiac resynchronization therapy, vol. 66, no. 2, pp. 343–353
49. Murugesan B et al (2018) Ecgnet: Deep network for arrhythmia classification, In 2018 IEEE International Symposium on Medical Measurements and Applications (MeMeA), pp. 1–6: IEEE
50. Salem M, Taheri S, Yuan JS (2018) ECG arrhythmia classification using transfer learning from 2-dimensional deep CNN features, In 2018 IEEE Biomedical Circuits and Systems Conference (BioCAS), pp. 1–4: IEEE
51. Alquran H, Alqudah A, Abu-Qasmieh I, Al-Badarneh A, Almashaqbeh SJNNW (2019) ECG classification using higher order spectral estimation and deep learning techniques, vol. 29, no. 4, pp. 207–219
52. Byeon Y-H, Pan S-B, Kwak K-CJS (2019) Intelligent deep models based on scalograms of electrocardiogram signals for biometrics, vol. 19, no. 4, p. 935
53. Tadesse GA et al (2019) Cardiovascular disease diagnosis using cross-domain transfer learning, In 2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), pp. 4262–4265: IEEE
54. Diker A, Cömert Z, Avcı E, Toğaçar M, Ergen B (2019) A Novel Application based on Spectrogram and Convolutional Neural Network for ECG Classification," In 2019 1st International Informatics and Software Engineering Conference (UBMYK), pp. 1–6: IEEE
55. Cao X-C, Yao B, Chen B-QJIA (2019) Atrial fibrillation detection using an improved multi-Scale decomposition enhanced residual convolutional neural network, vol. 7, pp. 89152–89161
56. Jiang F et al (2019) A Transfer Learning Approach to Detect Paroxysmal Atrial Fibrillation Automatically Based on Ballistocardiogram Signal, vol. 9, no. 9, pp. 1943–1949
57. Van Steenkiste G, van Loon G, Crevecoeur G. JSR (2020) Transfer Learning in ECG Classification from Human to Horse Using a Novel Parallel Neural Network Architecture, vol. 10, no. 1, pp. 1–12
58. Mazo C, Bernal J, Trujillo M, Alegre E (2018) Transfer learning for classification of cardiovascular tissues in histological images. Comput Methods Prog Biomed 165:69–76
59. Dietlmeier J, McGuinness K, Rugonyi S, Wilson T, Nuttall A, O'Connor NEJPRL (2019) Few-shot hypercolumn-based mitochondria segmentation in cardiac and outer hair cells in focused ion beam-scanning electron microscopy (FIB-SEM) data, vol. 128, pp. 521–528
60. Miyagawa M, Costa MGF, Gutierrez MA, Costa JPGF, Filho CFFJIAC (2019) Detecting Vascular Bifurcation in IVOCT Images Using Convolutional Neural Networks With Transfer Learning, vol. 7, pp. 66167–66175
61. Blanquer I, Brasileiro F, Brito A, Calatrava A, Carvalho A, Fetzer C, Figueiredo F, Guimarães RP, Marinho L, Meira W Jr, Silva A, Alberich-Bayarri Á, Camacho-Ramos E, Jimenez-Pastor A, Ribeiro ALL, Nascimento BR, Silva F (Sep 2020) Federated and secure cloud services for building medical image classifiers on an intercontinental infrastructure. Future Generation Comput Syst Int J Esci 110:119–134
62. Vu CC, Siddiqui ZA, Zamdborg L, Thompson AB, Quinn TJ, Castillo E, Guerrero TM (2020) Deep convolutional neural networks for automatic segmentation of thoracic organs-at-risk in radiation oncology - use of non-domain transfer learning. J Appl Clin Med Phys 21(6):108–113
63. Huynh BQ, Li H, Giger MLJJOMI (2016) Digital mammographic tumor classification using transfer learning from deep convolutional neural networks, vol. 3, no. 3, p. 034501
64. Kandaswamy C, Silva LM, Alexandre LA, Santos JMJJOBS (2016) High-content analysis of breast cancer using single-cell deep transfer learning, vol. 21, no. 3, pp. 252–259
65. Samala RK, Chan HP, Hadjiiski L, Helvie MA, Wei J, Cha KJMP (2016) Mass detection in digital breast tomosynthesis: Deep

convolutional neural network with transfer learning from mammography, vol. 43, no. 12, pp. 6654–6666
66. Dhungel N, Carneiro G, Bradley APJMIA (2017) A deep learning approach for the analysis of masses in mammograms with minimal user intervention, vol. 37, pp. 114–128
67. Kooi T, van Ginneken B, Karssemeijer N, den Heeten AJMP (2017) Discriminating solitary cysts from soft tissue lesions in mammography using a pretrained deep convolutional neural network, vol. 44, no. 3, pp. 1017–1027
68. Samala RK et al (2017) Multi-task transfer learning deep convolutional neural network: application to computer-aided diagnosis of breast cancer on mammograms, vol. 62, no. 23, p. 8894
69. Yap MH et al (2017) Automated breast ultrasound lesions detection using convolutional neural networks, vol. 22, no. 4, pp. 1218–1226
70. Chougrad H, Zouaki H, Alheyane OJCM (2018) Deep convolutional neural networks for breast cancer screening. P. I. Biomedicine 157:19–30
71. Mohamed AA, Berg WA, Peng H, Luo Y, Jankowitz RC, Wu SJMP (2018) A deep learning method for classifying mammographic breast density categories, vol. 45, no. 1, pp. 314–321
72. Samala RK, Chan H-P, Hadjiiski L, Helvie MA, Richter CD, Cha KHJITOMI (2018) Breast cancer diagnosis in digital breast tomosynthesis: effects of training sample size on multi-stage transfer learning using deep neural nets, vol. 38, no. 3, pp. 686–696
73. Samala RK et al (2018) Evolutionary pruning of transfer learned deep convolutional neural network for breast cancer diagnosis in digital breast tomosynthesis, vol. 63, no. 9, p. 095005
74. Zhang J, Chen B, Zhou M, Lan H, Gao FJIA (2018) Photoacoustic image classification and segmentation of breast cancer: A feasibility study, vol. 7, pp. 5457–5466
75. Byra M et al (2019) Breast mass classification in sonography with transfer learning using a deep convolutional neural network and color conversion, vol. 46, no. 2, pp. 746–755
76. Khan S, Islam N, Jan Z, Din I. U, Rodrigues JJCJPRL (2019) A novel deep learning based framework for the detection and classification of breast cancer using transfer learning, vol. 125, pp. 1–6
77. Mendel K, Li H, Sheth D, Giger MJAR (2019) Transfer learning from convolutional neural networks for computer-aided diagnosis: a comparison of digital breast tomosynthesis and full-field digital mammography, vol. 26, no. 6, pp. 735–743
78. Xie J, Liu R, Luttrell IV J, Zhang CJFIG (2019) Deep learning based analysis of histopathological images of breast cancer, vol. 10, p. 80
79. Yu S, Liu L, Wang Z, Dai G, Xie YJSCTS (2019) Transferring deep neural networks for the differentiation of mammographic breast lesions, vol. 62, no. 3, pp. 441–447
80. Zhu Z et al (2019) Deep learning for identifying radiogenomic associations in breast cancer, vol. 109, pp. 85–90
81. Zhu Z et al (2019) Deep learning analysis of breast MRIs for prediction of occult invasive disease in ductal carcinoma in situ, vol. 115, p. 103498
82. Chaves E, Goncalves CB, Albertini MK, Lee S, Jeon G, Fernandes HC (Jun 2020) Evaluation of transfer learning of pretrained CNNs applied to breast cancer detection on infrared images. Appl Opt 59(17):E23–E28
83. Chen P, Chen Y, Deng Y, Wang Y, He P, Lv X, Yu J (Aug 2020) A preliminary study to quantitatively evaluate the development of maturation degree for fetal lung based on transfer learning deep model from ultrasound images. Int J Comput Assist Radiol Surg 15(8):1407–1415
84. Chougrad H, Zouaki H, Alheyane O (2020) Multi-label transfer learning for the early diagnosis of breast cancer. Neurocomputing 392:168–180
85. Hu QY, Whitney HM, Giger ML (2020) A deep learning methodology for improved breast cancer diagnosis using multiparametric MRI. Sci Rep 10(1)
86. Sawada Y, Kozuka K (2015) Transfer learning method using multi-prediction deep boltzmann machines for a small scale dataset, In 2015 14th IAPR International Conference on Machine Vision Applications (MVA), pp. 110–113: IEEE
87. Shouno H, Suzuki S, Kido S (2015) A transfer learning method with deep convolutional neural network for diffuse lung disease classification, In International Conference on Neural Information Processing, pp. 199–207: Springer
88. Christodoulidis S, Anthimopoulos M, Ebner L, Christe A, Mougiakakou SJIJOB (2016) H Informatics, Multisource transfer learning with convolutional neural networks for lung pattern analysis, vol. 21, no. 1, pp. 76–84
89. Paul R et al (2016) Deep feature transfer learning in combination with traditional features predicts survival among patients with lung adenocarcinoma, vol. 2, no. 4, p. 388
90. Seelan LJ, Suresh LP, Veni SK (2016) Automatic extraction of Lung lesion by using optimized toboggan based approach with feature normalization and transfer learning methods, In 2016 International Conference on Emerging Technological Trends (ICETT), pp. 1–10: IEEE
91. Shen W et al (2016) Learning from experts: Developing transferable deep features for patient-level lung cancer prediction, In International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 124–131: Springer
92. Nibali A, He Z, Wollersheim DJIJOCAR (2017) Pulmonary nodule classification with deep residual networks. Surgery 12(10): 1799–1808
93. Hussein S, Cao K, Song Q, Bagci U (2017) Risk stratification of lung nodules using 3D CNN-based multi-task learning, In International conference on information processing in medical imaging, pp. 249–260: Springer
94. Shan H, Wang G, Kalra MK, de Souza R, Zhang J (2017) Enhancing transferability of features from pretrained deep neural networks for lung nodule classification, In Proceedings of the 2017 International Conference on Fully Three-Dimensional Image Reconstruction in Radiology and Nuclear Medicine
95. Wang C, Elazab A, Wu J, Hu QJCMI (2017) Lung nodule classification using deep feature fusion in chest radiography. Graphics 57:10–18
96. da Nóbrega RVM, Peixoto SA, da Silva SPP, Rebouças Filho PP (2018) Lung nodule classification via deep transfer learning in CT lung images, In 2018 IEEE 31st International Symposium on Computer-Based Medical Systems (CBMS), pp. 244–249: IEEE
97. Hosny A et al (2018) Deep learning for lung cancer prognostication: A retrospective multi-cohort radiomics study, vol. 15, no. 11, p. e1002711
98. Dey R, Lu Z, Hong Y (2018) Diagnostic classification of lung nodules using 3D neural networks, In 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018), pp. 774–778: IEEE
99. Fang T (2018) A novel computer-aided lung cancer detection method based on transfer learning from GoogLeNet and median intensity projections, In 2018 IEEE International Conference on Computer and Communication Engineering Technology (CCET), pp. 286–290: IEEE
100. Nishio M et al (2018) Computer-aided diagnosis of lung nodule classification between benign nodule, primary lung cancer, and metastatic lung cancer at different image size using deep convolutional neural network with transfer learning, vol. 13, no. 7
101. Hussein S, Kandel P, Bolan CW, Wallace MB, Bagci UJITOMI (2019) Lung and pancreatic tumor characterization in the deep learning era: novel supervised and unsupervised learning approaches, vol. 38, no. 8, pp. 1777–1787

102. Lakshmi D, Thanaraj KP, Arunmozhi MJIJOIS (2019) Technology Convolutional neural network in the detection of lung carcinoma using transfer learning approach
103. Li Y, Zhang L, Chen H, Yang NJIA (2019) Lung nodule detection with deep learning in 3D thoracic MR images, vol. 7, pp. 37822–37832
104. Shi Z et al (2019) A deep CNN based transfer learning method for false positive reduction, vol. 78, no. 1, pp. 1017–1033
105. Zhang S et al (2019) Computer-aided diagnosis (CAD) of pulmonary nodule of thoracic CT image using transfer learning, vol. 32, no. 6, pp. 995–1007
106. Huang X, Lei Q, Xie T, Zhang Y, Hu Z, Zhou QJAPA (2020) Deep Transfer Convolutional Neural Network and Extreme Learning Machine for Lung Nodule Diagnosis on CT images
107. Wankhade NV, Patey MA (2013) Transfer learning approach for learning of unstructured data from structured data in medical domain, In 2013 2nd International Conference on Information Management in the Knowledge Economy, pp. 86–91: IEEE
108. Marsh JN et al (2018) Deep learning global glomerulosclerosis in transplant kidney frozen sections, vol. 37, no. 12, pp. 2718–2728
109. Zheng Q, Tastan G, Fan Y (2018) Transfer learning for diagnosis of congenital abnormalities of the kidney and urinary tract in children based on Ultrasound imaging data, In 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018), pp. 1487–1490: IEEE
110. Zheng Q, Furth SL, Tasian GE, Fan YJJOPU (2019) Computer-aided diagnosis of congenital abnormalities of the kidney and urinary tract in children based on ultrasound imaging data by integrating texture image features and deep transfer learning image features, vol. 15, no. 1, pp. 75. e1–75. e7
111. Efremova DB, Konovalov DA, Siriapisith T, Kusakunniran W, Haddawy PJAPA (2019) Automatic segmentation of kidney and liver tumors in CT images
112. Hao P-Y et al (2019) Texture branch network for chronic kidney disease screening based on ultrasound images, pp. 1–10
113. Kannan S et al (2019) Segmentation of glomeruli within trichrome images using deep learning, vol. 4, no. 7, pp. 955–962
114. Kuo C-C et al (2019) Automation of the kidney function prediction and classification through ultrasound-based kidney imaging using deep learning, vol. 2, no. 1, pp. 1–9
115. Wu Z et al (2019) PASnet: A Joint Convolutional Neural Network for Noninvasive Renal Ultrasound Pathology Assessment, In 2019 IEEE 7th International Conference on Bioinformatics and Computational Biology (ICBCB), pp. 96–99: IEEE
116. Yin S et al (2018) Subsequent boundary distance regression and pixelwise classification networks for automatic kidney segmentation in ultrasound images
117. Yin S et al (2020) Automatic kidney segmentation in ultrasound images using subsequent boundary distance regression and pixelwise classification networks, vol. 60, p. 101602
118. Ayyar M, Mathur P, Shah RR, Sharma SG (2018) Harnessing ai for kidney glomeruli classification, In 2018 IEEE International Symposium on Multimedia (ISM), pp. 17–20: IEEE
119. Mathur P, Ayyar M, Shah RR, Sharma S (2019) Exploring Classification of Histological Disease Biomarkers from Renal Biopsy Images, In 2019 IEEE Winter Conference on Applications of Computer Vision (WACV), pp. 81–90: IEEE
120. Huang C, Lu Y, Lan Y, Chen S, Guo S, Zhang G (2020) Automatic segmentation of bioabsorbable vascular stents in intravascular optical coherence images using weakly supervised attention network, Futur Gener Comput Syst, 2020/07/27/
121. Huang C et al (2020) A Deep Segmentation Network of Multi-scale Feature Fusion based on Attention Mechanism for IVOCT Lumen Contour, IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol. PP, pp. 1–1, 02/14
122. Huang C et al (2019) A new pulse coupled neural network (PCNN) for brain medical image fusion empowered by shuffled frog leaping algorithm. Front Neurosci 13:03/20
123. Huang C et al (2020) A New Transfer Function for Volume Visualization of Aortic Stent and Its Application to Virtual Endoscopy. J ACM Trans Multimedia Comput Commun Appl 16(2s %):Article 65
124. Huang C et al (2019) Patient-Specific Coronary Artery 3D Printing Based on Intravascular Optical Coherence Tomography and Coronary Angiography. Complexity 2019:1–10, 12/23
125. Huang C et al (2018) A New Framework for the Integrative Analytics of Intravascular Ultrasound and Optical Coherence Tomography Images, IEEE Access, vol. PP, pp. 1–1, 05/22
126. da Nóbrega RVM, Reboucas PP, Rodrigues MB, da Silva SPP, Dourado C, de Albuquerque VHC (2020) Lung nodule malignancy classification in chest computed tomography images using transfer learning and convolutional neural networks. Neural Comput Applic 32(15):11065–11082
127. Lin F et al (2020) A CT-based deep learning model for predicting the nuclear grade of clear cell renal cell carcinoma. Eur J Radiol 129 Art. no. 109079

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.