CNN Based Image Classification

Xiaoling Long
SIST
ShanghaiTech Univ.
longxl@shanghaitech.edu.cn

Hongyu Chen
SIST
ShanghaiTech Univ.
chenhy3@shanghaitech.edu.cn

Ruiqi Luo
SIST
ShanghaiTech Univ.
luorq@shanghaitech.edu.cn

Abstract

The convolution networks make a great success in image classification since AlexNet proposed for classification in ImageNet chanllenge. This End-to-End training method make researcher focus more on how to build an efficient network. The user focus more on how to use convolution networks in their application. After AlexNet, there are many ConvNet architecture proposed, such as VGGNet and ResNet. These ConvNet architectures are widely used in many actual real applications.

1. Introduction

Image classification is a fundamental challenge in computer vision [7]. Consider the problem of detecting objects from a category, such as people or cars, in static images. This is a difficult problem because objects in each category can vary greatly in appearance. Variations arise from changes in illumination, viewpoint, and intra-class variability of shape and other visual properties among object instances. For example, people wear different clothes and take a variety of poses while cars come in various shapes and colors.Classification between the objects is easy task for humans but it has proved to be a complex problem for machines.Image classification refers to the labeling of images into one of a number of predefined categories. Image classification is an important and challenging task in various application domains, including biomedical imaging, biometry, video surveillance, vehicle navigation, industrial visual inspection, robot navigation, and remote sensing [12].

Image classification uses both supervised and unsupervised Traditional methods of computer vision and machine learning cannot match human performance on tasks such as the recognition of handwritten digits or traffic signs. [4]

Nonparametric classifiers such as decision tree classifier, neural network, SVM classifier and knowledge based classification are also very common in image classification.

Disadvantages about the deep learning.For the deep learning based algorithm required the ability of labeled samples for training. The collection of labeled data is a time consuming process as well as costly.

2. Evaluation standard

Two method are commons used for the image classification one is TOP 1 error another is TOP 5 error. TOP 1 error: The correct answer is the top guess.TOP 5 error: The correct answer is in the top-5 guess

3. Dataset

There are several common dataset for the image classification.such as the MINIST dataset [17],ImageNet [28] COCO dataset [22].PASCAL [6] CIFAR [14]and the Open Image which is from google.

MNIST is a hand-written digital database with 60,000 training sample sets and 10,000 test sample sets. Each sample image has a height of 28*28.This dataset is stored in binary and cannot be viewed directly in image format, but it is easy to find tools to convert it to image format.

ImageNet is an image dataset organized according to the WordNet hierarchy. Each meaningful concept in WordNet, possibly described by multiple words or word phrases, is called a ”synonym set” or ”synset”. There are more than 100,000 synsets in WordNet, majority of them are nouns (80,000+).

COCO is a large-scale object detection, segmentation, and captioning dataset. COCO has several features:Object segmentation,Recognition in context Superpixel stuff segmentation, 330K images (>200K labeled), 1.5 million object instances, 80 object categories, 91 stuff categories, 5 captions per image, 250,000 people with keypoints.

The main goal of this challenge is to recognize objects from a number of visual object classes in realistic scenes (i.e. not pre-segmented objects). It is fundamentally a supervised learning learning problem in that a training set of labelled images is provided. The twenty object classes that have been selected are:

Typical CIFAR daset contains the CIFAR-10 dataset ,he CIFAR-10 dataset consists of 60000 32x32 colour images in 10 classes, with 6000 images per class. There are 50000 training images and 10000 test images.

The dataset is divided into five training batches and one test batch, each with 10000 images. The test batch contains exactly 1000 randomly-selected images from each class. The training batches contain the remaining images in random order, but some training batches may contain more images from one class than another. Between them, the training batches contain exactly 5000 images from each class. CIFAR-100 dataset This dataset is just like the CIFAR-10, except it has 100 classes containing 600 images each. There are 500 training images and 100 testing images per class. The 100 classes in the CIFAR-100 are grouped into 20 superclasses. Each image comes with a "fine" label (the class to which it belongs) and a "coarse" label (the superclass to which it belongs).

Open Image is a data set that contains 9 million image URLs. The images are divided into more than 6000 categories by tag annotations. The tags in the dataset contain more real-life entities than ImageNet (1000), which is enough for us to train deep neural networks from scratch.

4. CNN Based Image classification

The ImageNet Large Scale Visual Recognition Challenge(ILSRC) is a benchmark in object category classification and detection on 1000-classes and millions of images. After AlexNet achieved hugu success in ILSRC-2012, there are various variations of AlexNet [15] and many other types of ConvNet for image classification. Since that, ConvNet is widely used for image classification. [1] illustrate beriefly what is ConvNet, the components of ConvNet, the activation function in ConvNet, from LeNet to ResNet bunch of successful ConvNet and some open issues on CNN based image classification. LeNet is first proposed CNN based image classifition method.

LeNet [16] [18] is first proposed CNN based image classifition method. After that, AlexNet [15] brought Convolutional neural network into ILSRC. In this implementation, it contains 8 layers 5 convolutional and 3 fully-conneted. The main features of this network's architecture are ReLU [24] as activation function, overlapping pooling and skills for reducing overfitting. Based on ReLU and overlapping pooling, the network err rate has more or less reduction, and ReLU network learn several times faster than other saturating acitvation function such as tanh neurons network. Overfitting is common issue for machine learning, it uses data augmentation and dropout to avoid overfit. Dropout is a skill to reduce argument or increase hypothesis. There are many discuss about this. As to data augmentation, it enlarge the dataset by cropping 224×224 patched from the original image as well as these patchs' horizontal refliction. At the end averaging the predications as final socre.

[5] explored the generalization ability of ConvNet features, releasing DeCAF. Traditional image classifition pipeline is extracting feature, building bag of feature then put into classifier. [39] propose a CNN based feature extractor. This is an unsurpervising learning ConvNet or in other words, input image is also the kind of ground truth. After feature extracting, the final result can classify by any classifier. This strategy is used as a tool for visualizing and undestanding how ConvNet works [38].ConvNet have an impressive classification performance. However there is no clear understanding for why this work. [38] propose a architecture for visualizing and undestanding how ConvNet works.

OverFeat [29] is a integrated recognition, localization and detection. This network uses CNN extract feature from image and then perform classification and localization and detection. Multi-scale classification brought up in [29] to increase accuracy.

"Networ In Network" [21] propsed a new deep network structure. Different from conventional convolution layer, it brings up a new Mlpconv layer. This Mlpconv layer consist of sliding multilayer perceptron(MLP) window. In stead of fully-connected layer at the top of network, global average pooling is used to produce the resulting vector fed directly into the softmax layer. Verified by experiments, this NIN structure indeed works well on some benchmark datasets, and global average pooling can be regarded as regularizer. THis glolbal average pooling has no parameter. This stratety is used widely afterwards. 1×1 convolution conception proposed in [21] is used in GoogLeNet for dimension reduction.

AlexNet make a great success in image classification. Afterwards many various Network appear. GoogLeNet [35] proposed by google is a new level of oganization in the form of the " Inception module". This is a multi-scale arcthitecture. With the limitation of computational resource, it perform a 1×1 convolution to dimension reduction. Auxiliary classifier is also a brilliant strategy.This smart design makes a great success in ILSRC-2014. At the same time, the widely used ConvNet architecture VGGNet [31] won the first place in *Classification* $+$ *Localization competition*. It adds the number of layers up to $16 - 19$. Instead of 7×7 convolution filter in [31], it uses 3×3 as convolution filter. After multiple layers, it can get similar effect as 7×7 one. This design significantly reduces the parameters, and then reduces the overfitting. It also means the number of layers significantly increases. Altering convolutional layers and poolint layer became a common used Network architecture.

As the depth of ConvNet increasing, training gets more and more difficult. The training of very deep network becomes a open issue in CNN. Highway Networks [32, 33] propose *information highways* which allow unimpeded information flow across several layers. The *transform gate* $T(x, W_T)$ and the *carry gate* $C(x, W_C)$ proposed for de-

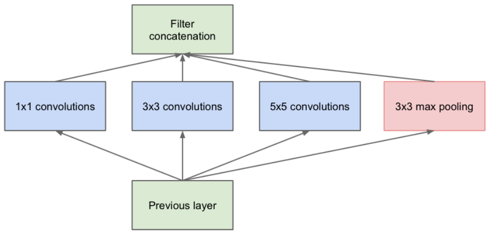


Figure 1. Original inception v1 network

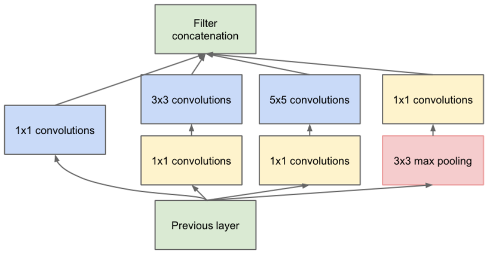


Figure 2. Advanced inception v1 network

cided how much flow pass through to output. The new model given by

$$y_{output} = H(x, W_H) \cdot T(x, W_T) + x \cdot C(x, W_C) \quad (1)$$

For simplicty, [9] set $C = 1 - T$. This design make training hundreds of layers be possible and the err rate just has slightly increase. This architecture promote the success of ResNet [9]. ResNet has similar structure as deep plain network stacked by dozens of convolution layers followed by global pooling layer and 1 fully-connected layer. except shortcup connection. This design has a residual representation which called deep residual learning. This archicture keeps parameter less than VGG-19 model even the network has 152 layers. This smart design make ResNet won first place in ILSRC-15.

Inception v1 network, stacking 1*1, 3*3, 5*5 conv and 3*3 pooling, on the one hand increases the network width, on the other hand increases the network adaptability to scale which is improved by [35]. The figure 1 is the most original version proposed in the paper. All the convolution kernels are done on all the outputs of the upper layer. The 5x5 convolution kernel requires too much calculation, resulting in The feature map is very thick. In order to avoid this phenomenon, the inception has the following structure. Before 3x3, before 5x5, add 1x1 convolutional kernels after max pooling to reduce the feature map thickness, which is the network structure of Inception v1.

Because Inception v1 has 5*5 kernels which means 25 parameters, so in this article [11], the author try to use two 3*3 kernels to replace the 5*5 kernel, which can reduce number of parameters and accelerate calculation. Inception v2 have batch normalization layer, which reduces internal covariate shift and normalized the output of each layer to an standard normal distribution.

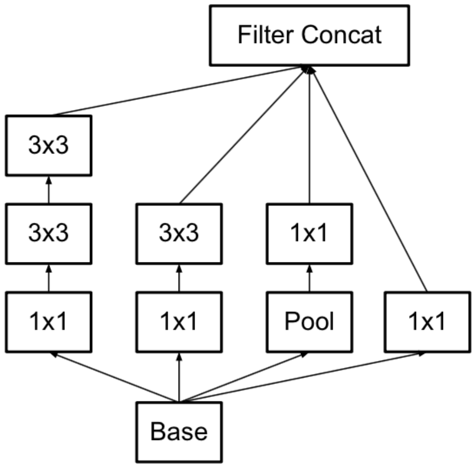


Figure 3. Difference between inception v2 and inception v1

One of the most important improvements to inception v3 is factorization, which decomposes 7x7 into two one-dimensional convolutions (1x7, 7x1) and so do 3x3 (1x3, 3x1). This method can speeding up calculations (so we can use excess calculations capability to deepen the network), and can also split a conv into two convs, which further increases the depth of the network and increases the nonlinearity of the network. It is also worth noting that the network input has changed from 224x224 to 299x299. 35x35/17x17/8x8 modules are designed. [36]

After Resnet v1 has been proposed, the google team find that the structure of Resnet can greatly speed up the training and improve performance at the same time. So they introduced the residual structure based on the inception v3, proposed the inception-resnet-v1 and inception-resnet-v2, and modified the inception module to propose an inception v4 structure [34]. The inception-v4 network has become deeper than v3. Before the GAP Inception-v3 includes four convolutional module operations (one regular convolution block and three inception structures), and Inception-v4 has a six convolutional module. Comparing the number of convolution kernels of both, Inception-v4 also has a lot more increase than Inception-v3.

Experiments based on inception v4 have found that similar results to the inception-resnet-v2 structure can be achieved without introducing a residual structure.So they combine the structure of Inception get an Inception-ResNet v2 network, and also design a deeper and more optimized Inception v4 model that can achieve performance comparable to Inception-ResNet v2.

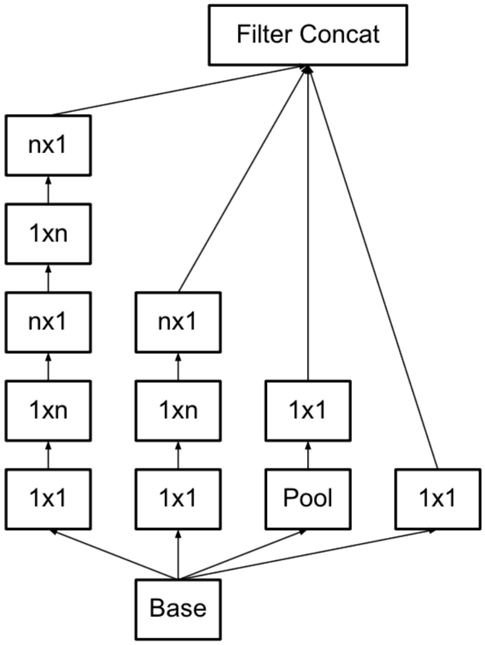


Figure 4. Difference between inception v3 and inception v2

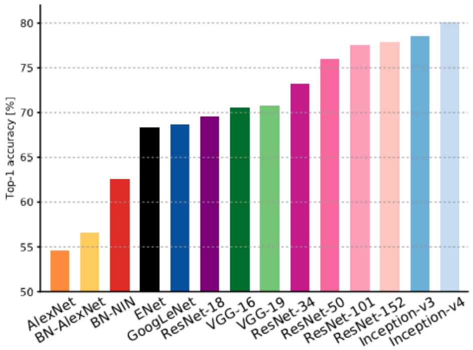


Figure 5. Single-crop top-1 validation accuracies for top scoring single-model architectures. We introduce with this chart our choice of colour scheme, which will be used throughout this publication to distinguish effectively different archi- tectures and their correspondent authors. Notice that networks of the same group share the same hue, for example ResNet are all variations of pink.

Figure5, in [3] shows the result that Top1 error vs the network. In the $x - axis$,we can see that the different types of the network. The accuracy get more and more accurate.The accuracy increase to about 80% from lower than 55%.

Top-1 one-crop accuracy versus amount of operations required for a single forward pass. The size of the blobs is proportional to the number of network pa- rameters; a legend is reported in the bottom right cor- ner, spanning fromparams. Both these figures share the same y-axis, and the grey dots highlight the centre of the blobs Figure6 shows the result that accuracy different size of the models. The larger the model, the big the radius.

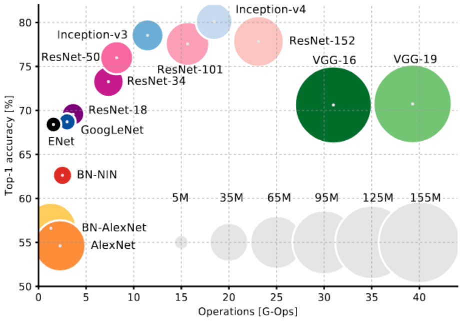


Figure 6. Top 1 vs operations ,size ∞ parameters

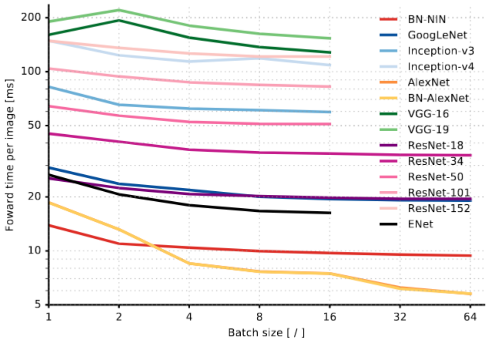


Figure 7. Top 1 vs operations ,size ∞ parameters

Figure7 in [3] shows the result that reference image. For each image that the VGG-NET need to take about $200ms$, This is because that the it is very large and there are a lot of parameters in it.

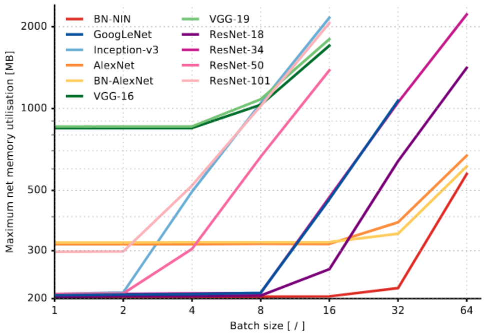


Figure 8. memory vs batch size

Figure in 8 in [3], shows the the memory vs the batch size ,from the image we can see different size of the memory need for each model. Different model need different memorize at the initial point.With the batch size increases the memory increases.

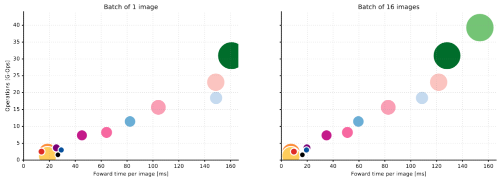


Figure 9. Operations vs inference time ,size∞ parameters.Relationship between operations and inference time, for batches of size 1 and 16 (biggest size for which all architectures can still run). Not surprisingly, we notice a linear trend, and therefore operations count represent a good estimation of inference time. Furthermore, we can notice an increase in the slope of the trend for larger batches, which correspond to shorter inference time due to batch processing optimisation.

Figure 9 in [3] shows the result that the operations with the inference time the operations is very large, the inference time is longer.

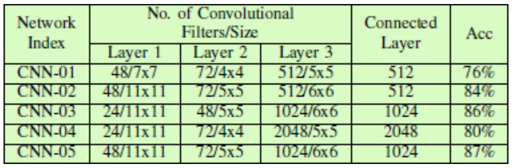
5. CNN based Medical Image Classification

5.1. Introduction

Convolutional neural networks (CNNs) have been used in the field of computer vision for decades. However, their true value had not been discovered until the ImageNet competition in 2012, a success that brought about a revolution through the efficient use of graphics processing units (GPUs), rectified linear units, new dropout regularization, and effective data augmentation. Acknowledged as one of the top 10 breakthroughs of 2013, CNNs have once again become a popular learning machine, now not only within the computer vision community but across various applications ranging from natural language processing to hyperspectral image processing and to medical image analysis. The main power of a CNN lies in its deep architecture, which allows for extracting a set of discriminating features at multiple levels of abstraction [37].

However, training a deep CNN with full training is very complicated. First, CNNs require a large amount of labeled training data. Second, training a deep CNN requires extensive computational and memory resources, without that the training process would be extremely time-consuming.Third, training a deep CNN is often complicated by overfitting. Therefore, deep learning from scratch can be tedious and time-consuming, demanding a great deal of diligence, patience, and expertise.

In this survey, I conducted an extensive set of experiments for 4 medical imaging applications: 1) polyp detection in colonoscopy videos [26] [40], 2) image quality assessment and classification in tissues and cells such as blood vessels videos [23] [20] [10] [13] [8] [2], 3) lung disease such as pulmonary embolism detection and so on in computed tomography (CT) images [30] [19],4) dental disease in X-ray image [25] and 5) intima-media boundary segmentation in ultrasonographic images [27].



Network Index	No. of Convolutional Filters/Size			Connected Layer	Acc
	Layer 1	Layer 2	Layer 3		
CNN-01	48/7x7	72/4x4	512/5x5	512	76%
CNN-02	48/11x11	72/5x5	512/6x6	512	84%
CNN-03	24/11x11	48/5x5	1024/6x6	1024	86%
CNN-04	24/11x11	72/4x4	2048/5x5	2048	80%
CNN-05	48/11x11	72/5x5	1024/6x6	1024	87%

Figure 10. Accuracy results from different CNN configurations

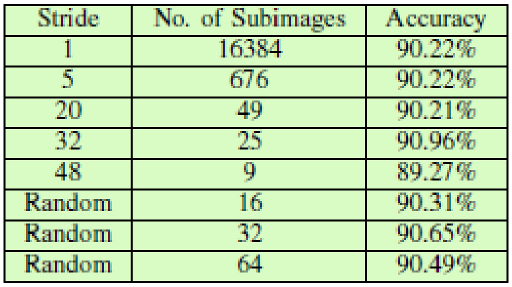
5.2. Polyp Detection

Colorectal cancer (CRC) is one of the leading causes of deathworldwide with about estimated 700 thousand deaths in 2012 [40]. Long-term follow-up studies confirmed that removal of adenomatous polyps reduces CRC mortality. Colonoscopy is the preferred technique for colon cancer screening and prevention. The goal of colonoscopy is to find and remove colonic polypsprecursors to colon cancer. But polyps can appear with substantial variations in color, shape, and size. The challenging appearance of polyps can often lead to misdetection [26]. Polyp miss-rates are estimated to be about 4% to 12%; however, a more recent clinical study is suggestive that this misdetection rate may be as high as 25%. So nowadays, there are many research groups start to use computer aided method such as CNN.

In the article [26], the author have a small dataset, which only have 100 images(75 abnormal images and 25 healthy images). After finishing the data augmentation which results in 800 images, they resized the 256*256 image to 128*128. In order to test the five architecture they established, he used cross validation method(56 for training and 6 for testing), the result can be seen in Figure 10, the accuracy is just 75% to 80%.

In order to improve the accuracy, in the evaluation phase, the author obtained the final decision for a 256*256 pixel image by majority voting of the decisions of all 128*128 pixel subimages(patches). This is a kind of fine-tuning. The redundancy of overlapping subimages can increase the system accuracy likewise to give the assurance of certainty for the overall decision. The result can be seen in figure 11. They also perform a random patch extraction and it can be concluded that there is not much difference between 16384 subimages or just 32 subimages (accuracy of 90.96%), saving considerable computation time and achieving good results.

In the second article [40], the author use a small datasets(PHW Database), this dataset consisted of 1104, 263 and 563 images without polyps, with hyperplasia polyps and adenomatous polyps, respectively, taken under either WL or NBI endoscopy. For fair comparison, 50 images from each class (nonpolyp, hyperplasia, and adenoma) were randomly selected as testing dataset, while the rest



Stride	No. of Subimages	Accuracy
1	16384	90.22%
5	676	90.22%
20	49	90.21%
32	25	90.96%
48	9	89.27%
Random	16	90.31%
Random	32	90.65%
Random	64	90.49%

Figure 11. Accuracy of different strides for overlapping subimages in the evaluation.

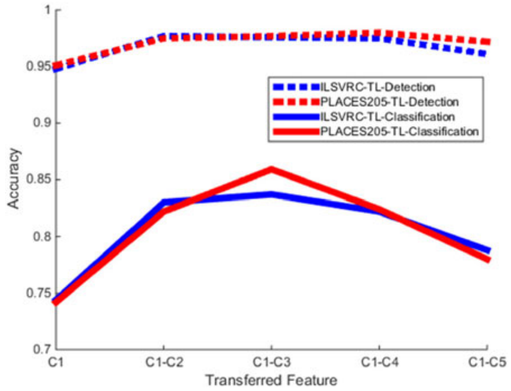


Figure 12. Average accuracy of the detection and classification tasks by transferring C1Cn features learned from ILSVRC and Places205 and using SVM as the classifier with a RBF kernel.

were treated as training dataset.

Because this dataset has an imbalanced number of images for each class, Previous study for polyp detection proposed to use an up/down sampling strategy to tackle such challenge. In this paper, the authors randomly down sampled the majority class to match the sample size of the minority class for both target tasks. The source dataset used ILSVRC and Places205 and trained for 450 000 iterations. They tested two tasks using this database, first is polyp detection and second polyp type classification. In order to do the evaluation, the authors used a feature engineering technique: bag-of-words for comparison. After finishing these tasks, we can see the results in the Figure 12,13,14. In these figures, we can see that transferring low-level CNN features gives better transfer learning performance for both target tasks and when a CNN structure is directly used for detection and classification. The performance of the proposed method is better in both tasks.

5.3. Tissue Detection and Classification

In this part, I will choose two typical article to discuss. In the first article [13], the author used feature vectors from several pre-trained structures, including networks

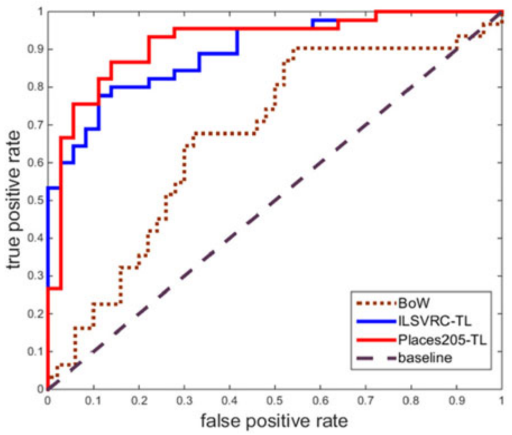


Figure 13. Typical ROC curve for polyp classification for PWH database.

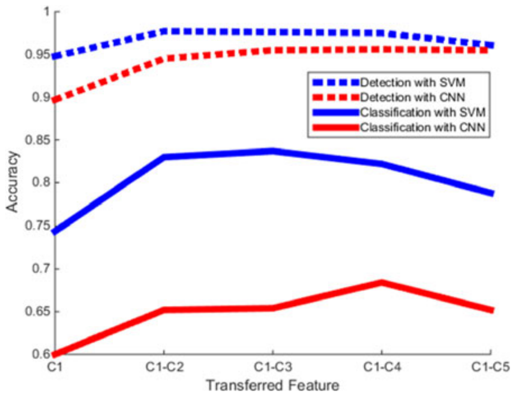


Figure 14. Average accuracy of the detection and classification tasks by transferring C1Cn features learned from ILSVRC and using either RBF kernel SVM or a fully connected CNN layer with a softmax classifier

with/without transfer learning to evaluate the performance of pre-trained deep features versus CNNs which have been trained by that specific dataset as well as the impact of transfer learning with a small number of samples. This experiment is done on Kimia Path24 dataset which consists of 27,055 histopathology training patches in 24 tissue texture classes along with 1,325 test patches for evaluation. In order to do this experiment, the author used fine-tuning method and a pre-trained CNN as a feature extractor and a fine-tuned CNN as a classifier.

The result shows in figure 15 that pre-trained networks are quite competitive against training from scratch. In this figure 15, VGG16 and CNN are quite similar, whereas the results for Inception-v3 are similar with the transfer-learned model outperforming the feature extractor. But considering Inception-v3 requires no extra effort and produces similar results with a linear SVM, one may prefer using it to training from scratch and fine-tuning a pre-trained net.

Scheme	Approach	η_p	η_w	η_{total}
Train from scratch	CNN_1 [17]	64.98%	64.75%	41.80%
Pre-trained features	FE-VGG16	65.21%	64.96%	42.36%
Fine-tuning the pre-trained net	TL-VGG16	63.85%	66.23%	42.29%
Pre-trained features	FE-Inception-v3	70.94%	71.24%	50.54%
Fine-tuning the pre-trained net	TL-Inception-v3	**74.87%**	**76.10%**	**56.98%**

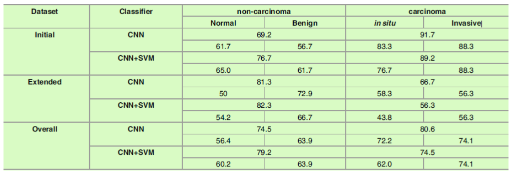
Figure 15. Comparing the results training form scratch , using deep features via a pre-trained network with no change (FE-VGG16), and classification after fine-tuning a pre-trained network (TL-VGG16, TL-Inception-v3). The best scores are highlighted in bold.(η_p means the patch-to-scan accuracy and η_n means whole-scan accuracy)

In the next article [2], the authors designed a specific CNN network which perform image-wise classification in four classes of medical relevance: normal tissue, benign lesion, in situ carcinoma and invasive carcinoma. The proposed CNN architecture is designed to integrate information from multiple histological scales, including nuclei, nuclei organization and overall structure organization. A data augmentation method is adopted to increase the number of cases in this training set. A SVM classification using the features extracted by the CNN is also used for comparison purposes.

The dataset is composed of an extended training set of 249 images, and a separate test set of 20 images. In these datasets, the four classes are balanced. The images were selected so that the pathology classification can be objectively determined from the image contents. An additional test set of 16 images is provided with images of increased ambiguity, which they denote as extended dataset.

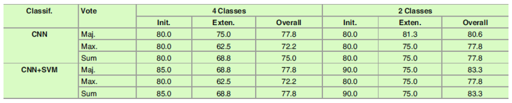
They first normalized the images. First, the colors of the images are converted to optical density (OD) using a logarithmic transformation. Then, they used singular value decomposition (SVD) to the OD tuples to find the 2D projections with higher variance. The resulting color space transform is then applied to the original image. Finally, the image histogram is stretched so that the dynamic range covers the lower 90% of the data.

Then they do two kinds of classification: Image-wise classification and CNN patch-wise classification. Image-wise classification first divided the origin image into twelve contiguous non-overlapping patches and then use one of three different patch patch methods: majority voting, maximum probability and sum of probabilities. CNN patch-wise classification used 75% of the data to do the training and validated on the remaining images.The validation set is randomly selected for each epoch. The training process stops after the stabilization of the validation accuracy with equal weight for all the classes (50 epochs). The authors also used the features extracted by the CNN to train a SVM classifier to do the comparison. The result can be seen in figure16 ,17and 18.



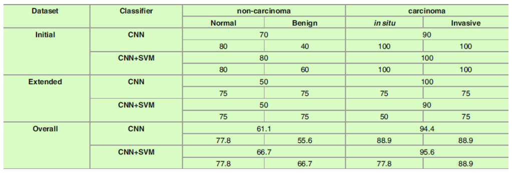
Dataset	Classifier	non-carcinoma		carcinoma	
		Normal	Benign	*in situ*	Invasive
Initial	CNN	69.2		91.7	
		61.7	56.7	83.3	88.3
	CNN+SVM	76.7		89.2	
		65.0	61.7	76.7	88.3
Extended	CNN	81.3		66.7	
		50	72.9	58.3	56.3
	CNN+SVM	82.3		56.3	
		54.2	66.7	43.8	56.3
Overall	CNN	74.5		80.6	
		56.4	63.9	72.2	74.1
	CNN+SVM	79.2		74.5	
		60.2	63.9	62.0	74.1

Figure 16. Patch-wise sensitivity (%) (2 and 4 classes).



Classif.	Vote	4 Classes			2 Classes		
		Init.	Exten.	Overall	Init.	Exten.	Overall
CNN	Maj.	80.0	75.0	77.8	80.0	81.3	80.6
	Max.	80.0	62.5	72.2	80.0	75.0	77.8
	Sum	80.0	68.8	75.0	80.0	75.0	77.8
CNN+SVM	Maj.	85.0	68.8	77.8	90.0	75.0	83.3
	Max.	80.0	62.5	72.2	80.0	75.0	77.8
	Sum	85.0	68.8	77.8	90.0	75.0	83.3

Figure 17. Image-wise accuracy (%) using different voting rules (2 and 4 classes).



Dataset	Classifier	non-carcinoma		carcinoma	
		Normal	Benign	*in situ*	Invasive
Initial	CNN	70		90	
		80	40	100	100
	CNN+SVM	80		100	
		80	60	100	100
Extended	CNN	50		100	
		75	75	75	75
	CNN+SVM	50		90	
		75	75	50	75
Overall	CNN	61.1		94.4	
		77.8	55.6	88.9	88.9
	CNN+SVM	66.7		95.6	
		77.8	66.7	77.8	88.9

Figure 18. Image-wise sensitivity (%) using majority voting (2 and 4 classes).

In figure16, we can see the result similar between the CNN and CNN+SVM. But the performance of this network is lower for the extended dataset due to its increased complexity. In figure 17 and 18, we can see that CNN+SVM get the best result with the majority voting method. In comparison, CNN's performance is only better for the extended set using majority voting. In addition, we can see that maximum probability is the worst performing method in both methods, which means that this method is not suit in this case.

5.4. Some kinds of lung Diseases Classification

Lung cancer is notoriously aggressive with a low long-term survival rate. Quantitative analysis in lung nodules using thoracic Computed Tomography(CT) has been a central focus for early cancer diagnosis, where CT phenotype provides a powerful tool to comprehensively capture nodule characteristics. The importance of diagnostically classifying malignant and benign nodules using CT images is to facilitate radiologists for nodule staging assessment and individual therapeutic planning. [30]

In the first article [30], the authors used the LIDC-IDRI datasets, which has 1375 nodule pictures(1100 for training and 275 for testing). In order to improve the speed and accuracy, the authors introduced an Multi-scale Convolutional Neural Networks(MCNN) model to do the lung nodule diagnostic classification. This CNN model take multi-scale raw nodule patches and remove the need of any hand-crafted feature engineering work. This network can also deal with noisy data in nodule CT.

Because of the clinical fact that nodule sizes vary re-

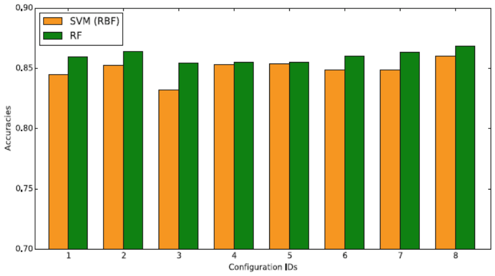
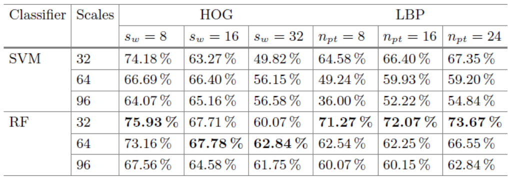


Figure 19. The classification performance of SVM with the RBF kernel and RF based on features from the MCNN using 8 different configurations. Each configuration is assigned to a unique ID for display convenience



Classifier	Scales	HOG			LBP		
		$s_w = 8$	$s_w = 16$	$s_w = 32$	$n_{pt} = 8$	$n_{pt} = 16$	$n_{pt} = 24$
SVM	32	74.18 %	63.27 %	49.82 %	64.58 %	66.40 %	67.35 %
	64	66.69 %	66.40 %	56.15 %	49.24 %	59.93 %	59.20 %
	96	64.07 %	65.16 %	56.58 %	36.00 %	52.22 %	54.84 %
RF	32	**75.93 %**	67.71 %	60.07 %	**71.27 %**	**72.07 %**	**73.67 %**
	64	73.16 %	**67.78 %**	**62.84 %**	62.54 %	62.25 %	66.55 %
	96	67.56 %	64.58 %	61.75 %	60.07 %	60.15 %	62.84 %

Figure 20. Performance using the HOG and LBP descriptors with different S_w and n_{pt}

markably, this network take patches from different scales(3 layers) as inputs in parallel. The parameter is shared between these layers to reduce parameter. When doing the evaluation task, the result is decided by all the layers. The authors use the HOG and uniform LBP descriptor and SVM and RF classifier to do the classification. The result can be seen in figure 19 and figure 20. In figure 20, the S_w means the size of the cell window for SVM and n_{pt} means the number of neighbourhood points for LBP.

The second article [19] is about using CNN to classify the ILD patterns. This experiment used an ILD database which contains 113 sets of HRCT images, with 2062 2D regions indicting the ILD category. In order to augment the dataset, the CT slices were divided into half-overlapping image and the only if 75% percent of its pixels falling inside the regions of interest will be adopted. The dataset thus contains 16220 image patches from 92 HRCT image sets, including 4348 norm patches, 1047 emphysema patches, 1953 ground glass patches, 2591 fibrosis patches, and 6281 micronodules patches.

The authors compared their classification results with three other feature extraction approaches: SIFT feature, LBP feature and unsupervised feature learning using RBM. The result can be seen in Figure 21 . In this figure21, we can see that their customized CNN method achieved the best classification preformance.

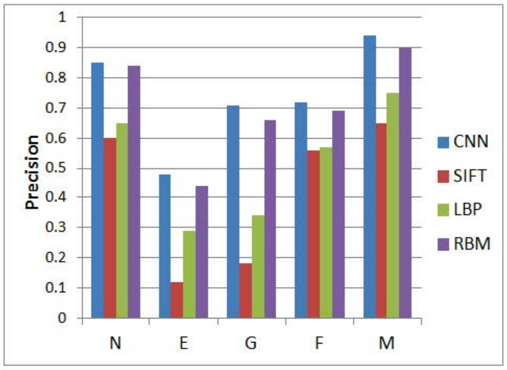
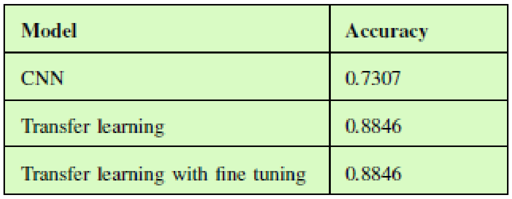


Figure 21. The classification results comparing proposed customized CNN method with SIFT, LBP and RBM



Model	Accuracy
CNN	0.7307
Transfer learning	0.8846
Transfer learning with fine tuning	0.8846

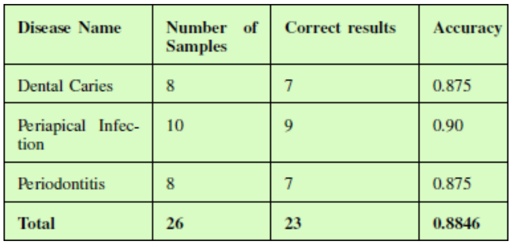
Figure 22. The comparison of different models

5.5. Dental Disease Classification in X-ray Images

The author found that there is no literature for dental disease classification, so the research group start to use CNN to deal with X-ray images and make some breakthrough. Orthopantomogram (OPG) and Radiovisiography (RVG) x-ray images are the most widely used tools for the diagnoses of dental diseases. Dental caries is one of the most common dental disease worldwide and it has different stages. So the CNN network in this experiment is used to classify mainly 3 classes (dental caries, periapical infection, periodontitis) [25].

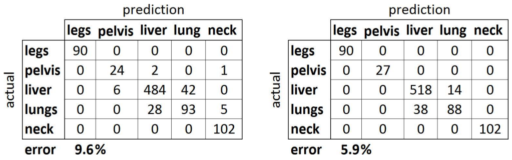
Because though the radiologists have large dataset of dental x-ray images, these x-ray images have individual privacy issues. So the dataset is very small in this experiment, just have 251 grey images of dimension 1000*1496. So the authors use transfer learning method to do the fine tuning and improve the accuracy very much. They changed some unfrozen layers used for training in order for the pre-trained model to be more adaptive to the training data.

They first resize these picture to 500*748, and then use 180 of 251 to do the training, 45 images for validation and 26 images for testing purpose. Because of the unavailability of the large dataset, CNN architecture could not perform well in this classification task. After they used transfer learning model to do the fine tuning, the accuracy is increased by 15.39% compared to pure CNN model, and achieved 88.46% accuracy, which is very encouraging.



Disease Name	Number of Samples	Correct results	Accuracy
Dental Caries	8	7	0.875
Periapical Infection	10	9	0.90
Periodontitis	8	7	0.875
Total	**26**	**23**	**0.8846**

Figure 23. Experimental results for transfer learning model



		prediction				
		legs	**pelvis**	**liver**	**lung**	**neck**
actual	**legs**	90	0	0	0	0
	pelvis	0	24	2	0	1
	liver	0	6	484	42	0
	lungs	0	0	28	93	5
	neck	0	0	0	0	102
	error	**9.6%**				

		prediction				
		legs	**pelvis**	**liver**	**lung**	**neck**
actual	**legs**	90	0	0	0	0
	pelvis	0	27	0	0	0
	liver	0	0	518	14	0
	lungs	0	0	38	88	0
	neck	0	0	0	0	102
	error	**5.9%**				

Figure 24. Confusion matrices on the original test images before and after data augmentation.

5.6. Intima-media Boundary Segmentation

Automated classification of human anatomy is an important prerequisite for many computer-aided diagnosis systems. The spatial complexity and variability of anatomy throughout the human body makes classification difficult. So the authors want to use CNN to do this classification. In this paper, the authors choose to use 4298 separate axial 2D key images to learn 5 anatomical classes(neck, lungs, liver, pelvis and legs) [27].

When applying the CNN to build the anatomy-specific classifier for CT images, because the authors want to classify these picture to 5 classes, so they choose 5 cascaded layers. All the convolutional filter kernel elements are trained from the data in a supervised fashion. In order to avoid overfitting, the fully-connected layers are constrained, using the *DropOut* method. The datasets are from the Picture Archiving and Communication System (PACS) of the Clinical Center of the National Institutes of Health. In order to enrich their data, they use spatial deformations to each image, using random translation, rotations and non-rigid deformations, which lead their datasets from hundred's picture to near 100 thousand pictures. Before import into the CNN, the author resize all the picture to 256*256 pixels.The authors use 80% of their total dataset to train the CNN and reserve 20% to do the test. After doing the experiments, the accuracy of this net can reach 94.1%, which can be seen in figure 24.This classification result is achieved in less than 1 minute on a modern desktop computer and GPU card (Dell Precision T7500, 24GB RAM, NVIDIA Titan Z).

5.7. Conclusion

In this part, I aimed to address to know how the CNN can be used on the medical image classification and the result these experiments made. My experiment, based on 4 distinct medical imaging applications from different imaging modality systems, have demonstrated that deep CNN are useful for medical image analysis. If the training data is limited, the fine-tuned CNN can perform better than fully trained CNN. I think the potential of CNNs for medical imaging applications is confirmed because both deeply fine-tuned CNNs and fully trained CNNs can outperform the corresponding handcrafted alternatives. We can also see that the speed is depend on the devices, the more powerful the graphics is, the quicker the CNN network use to train.

6. Conclusion

In this paper we have discussed about the different types of image classification techniques and many CNN based medical application. So this paper will help us in selecting an appropriate classification technique among all the available techniques.

References

[1] N. Aloysius and M. Geetha. A review on deep convolutional neural networks. In *Communication and Signal Processing (ICCSP), 2017 International Conference on*, pages 0588–0592. IEEE, 2017.

[2] T. Araújo, G. Aresta, E. Castro, J. Rouco, P. Aguiar, C. Eloy, A. Polónia, and A. Campilho. Classification of breast cancer histology images using convolutional neural networks. *PloS one*, 12(6):e0177544, 2017.

[3] A. Canziani, A. Paszke, and E. Culurciello. An analysis of deep neural network models for practical applications. *arXiv: Computer Vision and Pattern Recognition*, 2016.

[4] D. Ciregan, U. Meier, and J. Schmidhuber. Multi-column deep neural networks for image classification. *computer vision and pattern recognition*, pages 3642–3649, 2012.

[5] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, and T. Darrell. Decaf: A deep convolutional activation feature for generic visual recognition. In *International conference on machine learning*, pages 647–655, 2014.

[6] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results.

[7] P. F. Felzenszwalb, R. B. Girshick, D. A. Mcallester, and D. Ramanan. Visual object detection with deformable part models. *Communications of The ACM*, 56(9):97–105, 2013.

[8] O. Hadad, R. Bakalo, R. Ben-Ari, S. Hashoul, and G. Amit. Classification of breast lesions using cross-modal deep learning. In *Biomedical Imaging (ISBI 2017), 2017 IEEE 14th International Symposium on*, pages 109–112. IEEE, 2017.

[9] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[10] Y. Huang, H. Zheng, C. Liu, X. Ding, and G. K. Rohde. Epithelium-stroma classification via convolutional neural networks and unsupervised domain adaptation in histopathological images. *IEEE journal of biomedical and health informatics*, 21(6):1625–1632, 2017.

[11] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *international conference on machine learning*, pages 448–456, 2015.

[12] P. Kamavisdar, S. Saluja, and S. Agrawal. A survey on image classification approaches and techniques. *International Journal of Advanced Research in Computer and Communication Engineering*, 2(1):1005–1009, 2013.

[13] B. Kieffer, M. Babaie, S. Kalra, and H. Tizhoosh. Convolutional neural networks for histopathology image classification: Training vs. using pre-trained networks. *arXiv preprint arXiv:1710.05726*, 2017.

[14] A. Krizhevsky. Learning multiple layers of features from tiny images. 2009.

[15] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[16] Y. LeCun, B. E. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. E. Hubbard, and L. D. Jackel. Handwritten digit recognition with a back-propagation network. In *Advances in neural information processing systems*, pages 396–404, 1990.

[17] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[18] Y. LeOu11, L. Jackal, L. Bottou, A. Brunet, C. Cortes, J. Denker, H. Drucker, I. Guyon, U. Miiller, E. Séckinger, et al. Comparison of learning algorithms for handwritten digit recognition.

[19] Q. Li, W. Cai, X. Wang, Y. Zhou, D. D. Feng, and M. Chen. Medical image classification with convolutional neural network. In *Control Automation Robotics & Vision (ICARCV), 2014 13th International Conference on*, pages 844–848. IEEE, 2014.

[20] X. Li, W. Li, X. Xu, and W. Hu. Cell classification using convolutional neural networks in medical hyperspectral imagery. In *Image, Vision and Computing (ICIVC), 2017 2nd International Conference on*, pages 501–504. IEEE, 2017.

[21] M. Lin, Q. Chen, and S. Yan. Network in network. *arXiv preprint arXiv:1312.4400*, 2013.

[22] T. Lin, M. Maire, S. J. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollar, and C. L. Zitnick. Microsoft coco: Common objects in context. *european conference on computer vision*, pages 740–755, 2014.

[23] S. McIlroy, Y. Kubo, T. Trappenberg, J. Toguri, and C. Lehmann. In vivo classification of inflammation in blood vessels with convolutional neural networks. In *Neural Networks (IJCNN), 2017 International Joint Conference on*, pages 3022–3027. IEEE, 2017.

[24] V. Nair and G. E. Hinton. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th international conference on machine learning (ICML-10)*, pages 807–814, 2010.

[25] S. A. Prajapati, R. Nagaraj, and S. Mitra. Classification of dental diseases using cnn and transfer learning.

[26] E. Ribeiro, A. Uhl, and M. Häfner. Colonic polyp classification with convolutional neural networks. In *Computer-Based Medical Systems (CBMS), 2016 IEEE 29th International Symposium on*, pages 253–258. IEEE, 2016.

[27] H. R. Roth, C. T. Lee, H.-C. Shin, A. Seff, L. Kim, J. Yao, L. Lu, and R. M. Summers. Anatomy-specific classification of medical images using deep convolutional nets. In *Biomedical Imaging (ISBI), 2015 IEEE 12th International Symposium on*, pages 101–104. IEEE, 2015.

[28] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. S. Bernstein, et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015.

[29] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun. Overfeat: Integrated recognition, localization and detection using convolutional networks. *arXiv preprint arXiv:1312.6229*, 2013.

[30] W. Shen, M. Zhou, F. Yang, C. Yang, and J. Tian. Multi-scale convolutional neural networks for lung nodule classification. In *International Conference on Information Processing in Medical Imaging*, pages 588–599. Springer, 2015.

[31] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[32] R. K. Srivastava, K. Greff, and J. Schmidhuber. Highway networks. *arXiv preprint arXiv:1505.00387*, 2015.

[33] R. K. Srivastava, K. Greff, and J. Schmidhuber. Training very deep networks. In *Advances in neural information processing systems*, pages 2377–2385, 2015.

[34] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. *national conference on artificial intelligence*, pages 4278–4284, 2016.

[35] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, et al. Going deeper with convolutions. Cvpr, 2015.

[36] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. Rethinking the inception architecture for computer vision. *computer vision and pattern recognition*, pages 2818–2826, 2016.

[37] N. Tajbakhsh, J. Y. Shin, S. R. Gurudu, R. T. Hurst, C. B. Kendall, M. B. Gotway, and J. Liang. Convolutional neural networks for medical image analysis: Full training or fine tuning? *IEEE transactions on medical imaging*, 35(5):1299–1312, 2016.

[38] M. D. Zeiler and R. Fergus. Visualizing and understanding convolutional networks. In *European conference on computer vision*, pages 818–833. Springer, 2014.

[39] M. D. Zeiler, G. W. Taylor, and R. Fergus. Adaptive deconvolutional networks for mid and high level feature learning. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 2018–2025. IEEE, 2011.

[40] R. Zhang, Y. Zheng, T. W. C. Mak, R. Yu, S. H. Wong, J. Y. Lau, and C. C. Poon. Automatic detection and classification of colorectal polyps by transferring low-level cnn features from nonmedical domain. *IEEE journal of biomedical and health informatics*, 21(1):41–47, 2017.