Yadav and Jadhav *J Big Data* (2019) 6:113
https://doi.org/10.1186/s40537-019-0276-2

Journal of Big Data

RESEARCH Open Access

Deep convolutional neural network based medical image classification for disease diagnosis

Samir S. Yadav[1*] and Shivajirao M. Jadhav[2]

*Correspondence:
ssyadav@dbatu.ac.in
[1] Dr. Babasaheb Ambedkar Technological University, Raigad, Lonere, India
Full list of author information is available at the end of the article

Abstract

Medical image classification plays an essential role in clinical treatment and teaching tasks. However, the traditional method has reached its ceiling on performance. Moreover, by using them, much time and effort need to be spent on extracting and selecting classification features. The deep neural network is an emerging machine learning method that has proven its potential for different classification tasks. Notably, the convolutional neural network dominates with the best results on varying image classification tasks. However, medical image datasets are hard to collect because it needs a lot of professional expertise to label them. Therefore, this paper researches how to apply the convolutional neural network (CNN) based algorithm on a chest X-ray dataset to classify pneumonia. Three techniques are evaluated through experiments. These are linear support vector machine classifier with local rotation and orientation free features, transfer learning on two convolutional neural network models: Visual Geometry Group i.e., VGG16 and InceptionV3, and a capsule network training from scratch. Data augmentation is a data preprocessing method applied to all three methods. The results of the experiments show that data augmentation generally is an effective way for all three algorithms to improve performance. Also, Transfer learning is a more useful classification method on a small dataset compared to a support vector machine with oriented fast and rotated binary (ORB) robust independent elementary features and capsule network. In transfer learning, retraining specific features on a new target dataset is essential to improve performance. And, the second important factor is a proper network complexity that matches the scale of the dataset.

Keywords: CNN, Transfer learning, Capsule network, ORB, SVM, Image classification

Introduction

Effectively classifying medical images play an essential role in aiding clinical care and treatment. For example, Analysis X-ray is the best approach to diagnose pneumonia [1] which causes about 50,000 people to die per year in the US [2], but classifying pneumonia from chest X-rays needs professional radiologists which is a rare and expensive resource for some regions.

The use of the traditional machine learning methods, such as support vector methods (SVMs), in medical image classification, began long ago. However, these methods

have the following disadvantages: the performance is far from the practical standard, and the developing of them is quite slow in recent years. Also, the feature extracting and selection are time-consuming and vary according to different objects [3]. The deep neural networks (DNN), especially the convolutional neural networks (CNNs), are widely used in changing image classification tasks and have achieved significant performance since 2012 [4]. Some research on medical image classification by CNN has achieved performances rivaling human experts. For example, CheXNet, a CNN with 121 layers trained on a dataset with more than 100,000 frontal-view chest X-rays (ChestX-ray 14), achieved a better performance than the average performance of four radiologists. Moreover, Kermany et al. [3] propose a transfer learning system to classify 108,309 Optical coherence tomography (OCT) images, and the weighted average error is equal to the average performance of 6 human experts.

The medical images are hard to collect, as the collecting and labeling of medical data confronted with both data privacy concerns and the requirement for time-consuming expert explanations. In the two general resolving directions, one is to collect more data, such as crowdsourcing [5] or digging into the existing clinical reports [6]. Another way is studying how to increase the performance of a small dataset, which is very important because the knowledge achieved from the research can migrate to the research on big datasets. In addition to this, the most significant published chest X-ray image dataset (ChestX-ray 14) is still far smaller than the biggest general image dataset-ImageNet which has reached 14,197,122 instances at 2010 [7, 8].

CNN-based methods have various strategies to increase the performance of image classification on small datasets: One method is data augmentation [9–12]. Wang and Perez [13] researched the effectiveness of data augmentation in image classification. The authors found the traditional transform-based data augmentation has better performance than generative adversarial network (GAN) and other neural network-based methods. Another method is transfer learning [3, 12, 14, 15]. Kermany et al. [3] achieved 92% accuracy on a small pneumonia X-rays image dataset by transfer learning. The third method is the capsule network. Sabour et al. [16] invented a new neural network structure-capsule network, which achieves state-of-the-art performance on the Modified National Institute of Standards and Technology (MNIST) database [17]. And, also the best performance on other small datasets. Afshar et al. [18] have utilized capsule network to detect brain tumors and got 86.56% accuracy.

However, some gaps are needing to be noticed. A limitation of Kermany's research is they use the InceptionV3 model and stop retrain the convolutional layer of InceptionV3 because of the overfitting. Therefore, other models and the effects of retraining the convolutional layer will be evaluated in this research. Moreover, Afshar et al. [18] did not compare the performance of capsule network with other methods. Therefore, the contributions of this report include:

- Performance comparison of three different classification methods: SVM classifier with oriented fast and rotated binary robust independent elementary features (ORB), transfer learning of VGG16 and InceptionV3, and training capsule network from scratch.

- An analysis of the effects of data augmentation, network complexity, fine-tuned convolutional layer, and other preventing overfitting mechanics on the classification of small chest X-ray dataset by transfer learning of CNN.

This article conducts four groups of experiments. The SVM with ORB runs on a standard Machine. The convolutional neural network (CNN) related analyses are all run on a virtual machine with an Nvidia Tesla K80 Graphic card in Google Cloud [19].

The remainder of the article ordered as follows: "Literature review" section reviews the related literature on medical image classification. "Experimental design" section describes the design of experiments. "Experimental results" section presents the result of the experiments, and "Discussion" section discusses the results. Finally, the conclusion is drawn, and the future work described, followed by references.

Literature review

Medical image classification is a sub-subject of image classification. Many techniques in image classification can also be used on it. Such as many image enhanced methods to enhance the discriminable features for classification [20]. However, as CNN is an end to end solution for image classification, it will learn the feature by itself. Therefore, the literature about how to select and enhance features in the medical image will not be reviewed. The review mainly focuses on the application of traditional methods and CNN based transfer learning. And, on the capsule network on medical image related paper to investigate what factors in those models are essential to the final result and the gaps they haven't included in their work.

ORB and SVM application on medical image classification

Paredes et al. [21] use small patches of medical images as local features and k-nearest neighbor (k-NN) to classify the categorization of the whole medical image, finally achieving start-of-art accuracy. Parveen and Sathik [22] researched to detect Pneumonia from X-rays. The authors extracted features by discrete wavelet transform (DWT), wavelet frame transform (WFT) moreover, wavelet packet transform (WPT) and used Fuzzy C-means to detect Pneumonia. Caicedo et al. [23] use scale-invariant feature transform (SIFT) as a local feature descriptor and use support vector machines (SVM) classifiers to classify medical images and get state-of-art precision at 67%. However, SIFT is a patent algorithm. Thus, Rublee et al. [24] propose a free, faster local feature descriptor-oriented fast and rotated binary robust independent elementary features (ORB), which has the same performance as SIFT and even better performance than SIFT under some condition. SVM is also a high-performance classification algorithm, widely used in different medical image classification tasks by other researchers, and achieves an excellent performance [25, 26]. Therefore, this report uses ORB and SVM as the representation of the traditional methods.

CNN on medical image classification

With the different CNN-based deep neural networks developed and achieved a significant result on ImageNet Challenger, which is the most significant image classification and segmentation challenge in the image analyzing field [27]. The CNN-based deep

neural system is widely used in the medical classification task. CNN is an excellent feature extractor, therefore utilizing it to classify medical images can avoid complicated and expensive feature engineering. Qing et al. [28] presented a customized CNN with shallow ConvLayer to classify image patches of lung disease. The authors also found that the system can be generalized to other medical image datasets. Moreover, in other research, it also found that CNN based system can be trained from big chest X-ray (CXR) film dataset and state-of-art with high accuracy and sensitivity results on their dataset, like Stanford Normal Radiology Diagnostic Dataset containing more than 400,000 CXR and a new CXR database (ChestX-ray8), which consist of 108,948 frontal-view CXR [29]. Moreover, using limited data makes it hard to train an adequate model. Therefore the transfer learning of CNN is wildly used in medical image classification tasks. Kermany et al. [3] use InceptionV3 with ImageNet trained weight and transfer learning on a medical image dataset containing 108,312 optical coherence tomography (OCT) images. They got an average accuracy of 96.6%, with a sensitivity of 97.8% and a specificity of 97.4%. The authors also compared the results with six human experts. Most of the experts got high sensitivity but low specificity, while the CNN-based system got high values on both sensitivity and specificity. Moreover, on the average weight error measure, the CNN-based system exceeds two human experts. The authors also verified their system on a small pneumonia dataset, including about five thousand images, and achieved an average accuracy of 92.8%, with a sensitivity of 93.2% and a specificity of 90.1%. This system finally may help in accelerating diagnosis and referral of patients and therefore introduce early treatment, resulting in an increased cure rate. Moreover, Vianna [30] also studied how to utilize transfer learning to build an X-ray image classification system that is the critical component of a computer-aided-diagnosis system. The authors found a fine-tuned transfer learning system with data augmentation effectively alleviate overfitting problem and yield a better result than two other models: training from scratch and a transfer learning model with only a retrained last classification layer.

Capsule neural network on medical image classification

As mentioned in the previous section, the CapsNet was invented in 2017 [16]. Therefore, the research about it is not as fruitful as CNN. However, there is still some research on applying them to the different datasets and varying fields due to its excellent feature—Equivariance. This means the spatial relationship of objects in an image is kept, and at the same time, the result does not impact the object's orientation and size. Afshar et al. [18] applied CapsNet to classifying brain tumors on Magnetic Resonance Imaging (MRI) images and got 86.56% prediction accuracy with a modified CapsNet that reduces the feature maps from the original 256 to 64.

Moreover, Tomas and Robertas [31] presented a CapsNet based solution to classify four types of breast tissue biopsies from breast cancer histology images. They achieved 87% accuracy with the same high sensitivity. Jimenez-Sanchez et al. [5] evaluated the CapsNet on medical image challenges. The authors selected a CNN with three layers of ConvLayer as the baseline and compared CapsNet's performance with LeNet and the baseline on four datasets, MNIST, Fashion-MNIST, mitosis detection (TUPAC16) and diabetic retinopathy detection (DIARETDB1), with three conditions: the partial subset of the dataset, the imbalanced subset of the dataset and data

augmentation. The final result shows CapsNet performed better than the other two networks in a small, imbalanced dataset. Beşer et al. [32] implemented a sign language recognizing system by CapsNet and achieved 94.2% validation accuracy. Moreover, some researchers studied internal mechanics by varying network structures under different conditions. Xi et al. [33] studied the impact of different network structures on a complex dataset CIFAR10. The authors choose the following options:

1. Increase the number of primary capsule layers.
2. Increase the capsule number in primary capsule layer.
3. Assemble multiple models and average the result.
4. Adjust the scaling factor of reconstruction loss.
5. Add more ConvLayer.
6. Evaluate other activation function.

Finally, the authors found more ConvLayers and more models assembled, which have more effect on improving the final accuracy. Moreover, also they achieved the highest result with a 7-model assembled CapsNet with a more ConvLayer than the original version of Sabour's. Furthermore, The CapsNet of Tomas and Robertas used to classify breast cancer increased the ConvLayer to five layers. On the other hand, Afshar et al. [18] also evaluated the different options of CapsNet. They fine-tuned the input size, number of feature maps, number of ConvLayers, capsule number in primary CapsLayer, dimension number in Primary Capsule, and the neuron number in reconstruction layers. The authors got the best results with a CapsNet having a 64×64 input image (original is 28×28) and fewer feature map, which reduces to 64 from the original 256. Also, the authors found that increasing the routing iteration number beyond three will not improve the performance on the four datasets: MNIST, Fashion-MNIST, The Street View House Numbers (SVHN) dataset, and Canadian Institute for Advanced Research 10 (CIFAR10) dataset. From the previous reviews, it can be seen that the traditional method (SVM with ORB feature), CNN based transfer learning, and Capsule network can all use on the medical image dataset. Just looking at the value of accuracy on different datasets, CNN based transfer learning looks have better performance than the other two methods. However, they have not been compared to the same dataset. Therefore, this paper will compare their performance on the same dataset-the pneumonia dataset.

Moreover, there are so many different options when fine-tuning the parameter of those methods. The traditional method has so many features and classifying algorithms which can be evaluated. They cannot be iterated in this paper due to the limited time. As the baseline, the traditional method choose ORB as the feature and linear SVM as the classifier. As the data augmentation is a data preprocessing method that can apply to all three methods, it also will be evaluated on the traditional method. For CNN-based transfer learning, the layers of retrained ConvLayer, the complexity of classification layers, the dropout rate has significant effects on the final result. Therefore, they will be evaluated by this research. Based on the same research, the critical fact in capsule network: the number of the feature map, the number of the capsules, and the channels of the capsule will also be evaluated in this report.

Table 1 The composition of chest X-ray dataset

	Training dataset	Testing dataset
Normal	1349 (25.7%)	234 (37.5%)
Bacteria	2538 (48.5%)	242 (38.7%)
Virus	1345 (25.7%)	148 (23.7%)
Total	5232 (100%)	624 (100%)

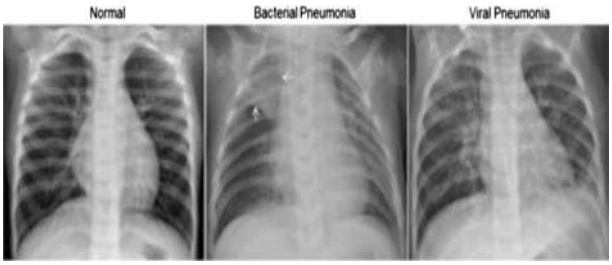

Fig. 1 Examples of chest X-rays [3]

Experimental design

Data neural network on medical image classification

The Dataset comes from the work of Kermnay et al. [34]. It contains two kinds of chest X-ray Images: NORMAL and PNEUMONIA, which are stored in two folders. In the PNEUMONIA folder, two types of specific PNEUMONIA can be recognized by the file name: BACTERIA and VIRUS. Table 1 describes the composition of the dataset. The training dataset contains 5232 X-ray images, while the testing dataset contains 624 images. In the training dataset, the image in the NORMAL class only occupies one-fourth of all data. In the testing dataset, the PNEUMONIA consists of 62.5% of all data, which means the accuracy of the testing data should higher 62.5%.

Figure 1 shows examples of chest X-rays from the dataset. The normal chest X-ray (left panel) depicts clear lungs without any areas of abnormal opacification in the image.

Bacterial pneumonia (middle) typically exhibits a focal lobar consolidation, in the right upper lobe (red rectangle), whereas viral pneumonia (right) manifests with a more diffuse interstitial pattern in both lungs.

Environment setup

Hardware

For ORB and SVM classification, an ordinary high-performance computer is enough, like 16G memory, i7 (2.3 GHz), and a 256G solid-state drive (SSD) disk. However, training a deep neural network should use GPU to accelerate the process. In this report, a Google Cloud GPU is used. A virtual machine instance with four core of CPU, 16G memory, and an NVIDIA Tesla K80 is used. Concerning the detail setup guide, please refer Google guide and other web pages [19, 35].

Software

To test the ORB and SVM classification, A python program which was initially used to classify plants are ported [36]. It was modified to use the new dataset and ran it on a laptop. An iteration of the test needs about four hours [15]. Because of the CNN-based method is computing intensively, so it needs to run on a VM in Google GPU Cloud. To test the Capsule network, a python capsule network implementation that aims to detect brain tumors was ported to the pneumonia dataset [37]. It also needs to be run on the GPU VM.

Data augmentation design

In this paper, three data augmentation algorithms evaluated. It can be seen from Table 2, Aug0 means using the original dataset without augmentation. Aug1 means simple geometrical transform of the image: such as randomly flip horizontally and vertically, randomly rotates within 0.05°, horizontal shear within the range 0.05 times the image width and zoom in within 0.05 times while Aug2 is a more complicated transform than Aug1. Besides all transforms of Aug1, it also does a slightly horizontal and veridical shift. To avoid the data exploration of the combination of the data augmentation models and classification algorithms, this paper only evaluates the effects of different augmentation algorithms on VGG16. This is the best classification algorithm for this paper. However, to analyze the effects of data augmentation, all three classification algorithms also evaluated on Aug0, and the best augmentation model got from this test.

ORB and SVM application experiments design

The ORB, VLAD, and SVM classification are chosen as the baseline. Two experiments conducted: First is classifying Normal and Pneumonia with the original dataset. Second does the same classification but with the best augmentation models.

Transfer learning experiments design

Because the chest X-ray dataset is small and different from ImageNet, whose weight used in the transfer learning experiments, therefore three group experiments conducted to fine-tune the final model. The first group of experiments aims to evaluate the effect of classification layer size on the final classification accuracy. Five models used on two CNN: VGG16 and InceptionV3 showed in Table 3. In the 2nd column,

Table 2 Augmentation models

Augmentation model	Augmentation parameters
Aug0	No Aug
Aug1	Rotation range = 0.05, shear range = 0.05, zoom range = 0.05, horizontal flip = True, vertical flip=True
Aug2	Rotation range = 3,width shift range = 0.05, height shift range = 0.05, shear range = 0.05, zoom range = 0.05, f fill mode = 'constant', cval = 0., horizontal flip = True, vertical flip = True

Table 3 Classification layer model configuration

	Configuration	VGG16 Training parameter	InceptionV3 Training parameter
Model1	GAPFC(4096) → FC(4096) → Softmax	18,890,754	25,182,210
Model2	GAP → Softmax	1026	4098
Model3	GAP → FC(512) → Dropout(0.5) → FC(256) → Dropout(0.5) → FC(128) → Dropout(0.5) → Softmax	427,138	1,213,570
Model4	GAP → FC(512) → Dropout(0.5) → Softmax	263,682	1,050,114
Model5	GAP → FC(512) → Dropout(0.5)→ FC(512) → Dropout(0.5) → FC(256) → Dropout(0.5) → Softmax	657,154	1,443,586

Table 4 Configuration of fine-tuned Convlayer model

	Configuration
ConvLayer Model	1 Best classification model with an unfrozen ConvLayer
ConvLayer Model2	Smaller classification model with an unfrozen ConvLayer
ConvLayer Model3	Better model in previous two model with two an unfrozen ConvLayer

the classification model described. Example, model3 consists of eight layers after the ConvLayers which are: global average pooling (GAP) layer, fully connected (FC) layer with 512 neurons, dropout layer with 50% drop rate, second FC layer with 256 neurons, second dropout layer with 50% drop rate, third FC layer with 128 neurons, third dropout layer with 50% drop rate and a classification layer with a SoftMax activation function. The last two columns list the parameters needed to be trained in VGG16 and InceptionV3. They used to indicate the complexity of the model.

The second group experiment aims to evaluate how many ConvLayers should be unfrozen and trained. A total of three experiments conducted (showed in Table 4). The first experiment evaluates the results of the best classification model with an unfrozen ConvLayer. Because the training parameter of the last ConvLayer is quite large, to prevent overfitting, the second experiment uses a smaller classification model. For testing the limit of the number of unfrozen ConvLayers, the third experiment unfreezes one more ConvLayers of the better ones in the previous two models.

The third group experiment fine-tunes other parameters based on the best model in the previous two experiment groups, such as increasing the drop rate of the dropout layer, reducing the learning rate, adding a batch normalization layer, which makes learning more stable and quicker.

Capsule neural network design

For CapsNet, the feature map number, the size of the PrimaryCaps layer and input image size impact on the performance of the classification. Thus, Table 5 shows the experiments aimed at evaluating the effects of those parameters.

Table 5 Experiments configuration of CapsNet

	Configuration
Test1	Aug0
Test2	Aug1
Test3	Aug1 with 64 feature maps and 64 input size
Test4	Aug1 with 64 feature maps and 128 input size
Test5	Aug1 with 32 feature maps and 64 input size
Test6	Aug1 with 32 feature maps and 128 input size
Test7	Aug1 with 32 feature maps and 48 input size
Test8	Aug1 with 24 feature maps and 64 input size
Test9	Aug1 with 16 feature maps and 64 input size
Test10	Aug1 with 32 feature maps, 64 input size and half primary capsule (4)
Test11	Aug1 with 32 feature maps, 64 input size, half primary capsule (4) and half capsule channel (16)
Test12	Aug1 with 32 feature maps, 64 input size, half primary capsule (4) and one fourth capsule channel (8)
Test13	Aug1 with 24 feature maps, 64 input size, half primary capsule (4) and half capsule channel (16)
Test14	Aug1 with 32 feature maps, 64 input size, half primary capsule (4) and half capsule channel (16) with more image by augmentation (10,000)

Table 6 Data augmentation experiments result

Aug algorithms	Total training images	VGG16 accuracy
Aug0	5232	0.882
Aug1	5232	0.898
Aug2	5232	0.895
Aug1	10,000	0.902
Aug2	10,000	0.879

Table 7 ORB And SVM classification experiments results

Augmentation	Accuracy
No Aug	0.74
Aug 20,000 images	0.776

Experimental results

In Table 6, the first column is the augmentation algorithms used in the test, the second column is the total training images generated by augmentation, and the last column is the average accuracy achieved by VGG16 transfer learning with all default parameter. From the result, it can be seen that the Aug1 is a better augmentation model than Aug2, therefore in the following experiments uses Aug1 as the default augmentation model.

ORB and SVM classification

In Table 7, the first column is the augmentation methods, and the second column is the average accuracy of the linear SVM classifier with ORB features. It can be seen the augmentation with more images increase the accuracy.

Table 8 Experimental result of evaluating classification model

	VGG16	Inception V3
Model1	0.881	0.629
Model2	0.631	0.818
Model3	0.898	0.875
Model4	0.873	0.857
Model5	0.885	0.869

Table 9 Experiments result of fine-tuned Convlayer

Model	VGG16
Model3 with last unfrozen ConvLayer	0.883
Model2 with last unfrozen ConvLayer	0.924
Model2 with last two unfrozen ConvLayer	0.9

Table 10 Experiments result of fine-tuned other parameters

Configuration	VGG16
Model2 with last unfrozen ConvLayer, lr 0.0009 and lr decay 0.8	0.9
Model2 with last unfrozen ConvLayer, lr 0.001 and lr decay 0.5	0.873
Model2 with last unfrozen ConvLayer and 20,000 augmentation image	0.871
Model2 with last unfrozen ConvLayer lr 0.0005, lr decay 0.5 and 10,000 augmentation image	0.902
Model3 with last unfrozen ConvLayer drop rate 0.7	0.885
Model3 with drop rate 0.7	0.906
Model3 with drop rate 0.7 and 20,000 augmentation image	0.922
Model3 with drop rate 0.7 and 30,000 augmentation image	0.922
Model2 with last unfrozen ConvLayer, batch normal layer, drop rate 0.5	0.912
Model2 with last unfrozen ConvLayer, batch normal layer, drop rate 0.5 and 20,000 augmentation image	0.906
Model2 with last unfrozen ConvLayer, batch normal layer, dropout 0.7, fc layer, dropout 0.5	0.916
Model2 with last unfrozen ConvLayer, batch normal layer, dropout 0.7, fc layer, dropout 0.5 and 20,000 augmentation image	0.875

Transfer learning classification

In Table 8, it can be seen that VGG16 is better than InceptionV3 and Model3 in VGG16 is the best classification model. Thus, in the following experiment, VGG16 and Model3 continue to be fine-tuned.

From Table 9, it can be seen that the classification model2 with the last unfrozen ConvLayer was the best model in all three experiments. Thus, the following experiments will continue to be fine-tuned.

The experiments in Table10 are explorational testing. The most successful models in previous experiments: Model3 and Model2 with the last unfrozen ConvLayer chosen as the baseline. Then according to the results and the effects of dropout, learning rate, and more training data, the parameters adjusted. In all the experiments, Model2 with the last unfrozen ConvLayer and all other default parameters still has the best results (showed in Tables 9, 10).

Table 11 Experiments result of CapsNet

Sr. no	Configuration	CapsNet
1	Aug0	0.748
2	Aug1	0.788
3	Aug1 with 64 feature maps and 64 input size	0.737
4	Aug1 with 64 feature maps and 128 input size	0.627
5	Aug1 with 32 feature maps and 64 input size	0.798
6	Aug1 with 32 feature maps and 128 input size	0.784
7	Aug1 with 32 feature maps and 48 input size	0.756
8	Aug1 with 24 feature maps and 64 input size	0.798
9	Aug1 with 16 feature maps and 64 input size	0.765
10	Aug1 with 32 feature maps, 64 input size and half primary capsule (4)	0.811
11	Aug1 with 32 feature maps, 64 input size, half primary capsule (4) and half capsule channel (16)	0.825
12	Aug1 with 32 feature maps, 64 input size, half primary capsule (4) and one fourth capsule channel (8)	0.752
13	Aug1 with 24 feature maps, 64 input size, half primary capsule (4) and half capsule channel (16)	0.825
14	Aug1 with 32 feature maps, 64 input size, half primary capsule (4) and half capsule channel (16) with more image by augmentation (10,000)	0.788

Capsule neural network

Table 11 shows the experiments processes and results by capsule network. The augmentation first is evaluated. The Aug1 is better than no augmentation, and the number of feature maps, input size, number of primary capsules, number of capsule channels, and number of training images vary. The best result comes from tests no. 11 and 13. They both have fewer feature maps, primary capsules, and capsule channels. The difference between them is the number of feature maps. One has 24 feature maps, while the other has 32 feature maps.

Verify on OCT dataset

To check if the findings can be used on other datasets, some experiments conducted on the OCT dataset which was published together with the Pneumonia dataset but included 108,309 OCT images. From Table 12, it can be seen, the best result obtained from test 5.

Discussion

The effects of data augmentation

Table 13 shows a summary of all the experiment test result between no augmentation and augmentation.

It can be seen that augmentation improves performance regardless of the model. That is because augmentation geometrically transforms the picture, which facilitates the machine learning algorithm to learn the underground feature without the impact of rotation and scale. However, from Table 6, it can be seen that complicated transforms are not always better than simple ones. Too complicated transforms introduce some noise in the feature that disturbs the learning process.

Table 12 Experiments result on OCT dataset

No.	Model	Accuracy
1	Model2 with last ConvLayer	0.934
2	Model3	0.828
3	Inceptionv3 Model2	0.791
4	Model2 with last two unfrozen ConvLayer	0.921
5	VGG16 with last ConvLayer → 4096 FC → 0.7 Dropout → 2048 FC → 0.5 dropout	0.954
6	VGG16 with last ConvLayer → 4096 FC → 0.7 Dropout → 2048 FC → 0.7 Dropout → 2048 FC → 0.5 dropout	0.937
7	VGG16 with last ConvLayer → 4096 FC → 0.8 dropout → 2048 FC → 0.7 dropout	0.938

Table 13 The comparison of data augmentation experiments

Model	Accuracy without augmentation	Accuracy with augmentation
ORB and SVM	0.74	0.776
VGG16	0.883	0.923
INV3	0.844	0.875
Caps Net	0.774	0.856

a

Model	VGG16 Params	InceptionV3 Params	VGG16	InceptionV3
Model2	1,026	4,098	0.63086	0.81836
Model4	263,682	1,050,114	0.87305	0.85742
Model3	427,138	1,213,570	0.89844	0.875
Model5	657,154	1,443,586	0.88477	0.86914
Model1	18,890,754	25,182,210	0.88086	0.62891

b

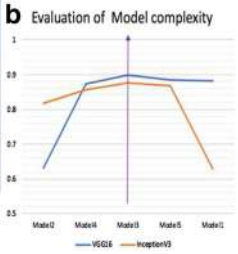

Fig. 2 The evaluation of model complexity: **a** Combinations of Tables 3 and 8, sorted in ascending order by the number of parameters. **b** The evaluation model complexity graph of the VGG16 and InceptionV3

The finding on fine-tune of transfer learning

1. Effects of model complicity of neural network: The left table of Fig. 2 is a combination of Tables 3 and 8 and sorted by the number of parameters in ascending order. It can be seen that the number of parameters has a significant impact on accuracy. Too many and too few parameters get poor results. The right graph of Fig. 2 shows that the highest results of VGG16 and InceptionV3 are in model3 that has the proper size of parameters that match the size of the database.

Table 14 Evaluation of dropout, batch normalization and learning rate for model2 with last unfrozen Convlayer

Learning rate	Decay rate	Training image	Dropout1	Dropout2	BNlayer	VGG16
0.001	0.9	20,000	0.5	NA	No	0.871
0.001	0.5	5323	0.5	NA	No	0.873
0.001	0.9	20,000	0.7	0.5	Yes	0.875
0.0009	0.8	5323	0.5	NA	No	0.9
0.0005	0.5	10,000	0.5	NA	No	0.902
0.001	0.9	20,000	0.7	NA	Yes	0.906
0.001	0.9	5323	0.5	NA	Yes	0.912
0.001	0.9	5323	0.7	0.5	Yes	0.916
0.001	0.9	5323	0.5	NA	No	0.924

Table 15 Evaluation of dropout, batch normalization and learning rate for model3

Model	Training image	Dropout1	VGG16
Model3 with last unfrozen Conv-Layer	5323	0.7	0.885
Model3	5323	0.5	0.898
Model3	5323	0.7	0.906
Model3	20,000	0.7	0.922
Model3	30,000	0.7	0.922

2. The effects of techniques to preventing overfitting: Table 14 shows the explorational test results of model2 with the last unfrozen ConvLayer. Because the whole training process tends to overfit, no single factor has a stable and significant impact on final accuracy. When comparing the result of model3 with different conditions (as in Table 15), it can be seen that the increasing dropout rate and augmentation number in each training iteration continually increase the accuracy. The opposite is the model with the last unfrozen ConvLayer. That is understandable because the last ConvLayer has too many parameters. Therefore, the training process is overfitting.

The finding on capsule network

1. The effects of feature maps: A series of experiments can unveil the effects of feature maps through fixing the input size (64), number of primary capsules (8), number of capsule channel (32) and varying the number of feature maps. The results in Fig. 3 show that the model with 24 and 32 feature maps got the best results.
2. The effects of input size: A series of experiments can unveil the effects of feature maps through fixing feature map size (32), number of primary capsules (8), number of capsule channel (32) and varying the input size. Figure 4 shows that the model with input size 64 got the best accuracy.

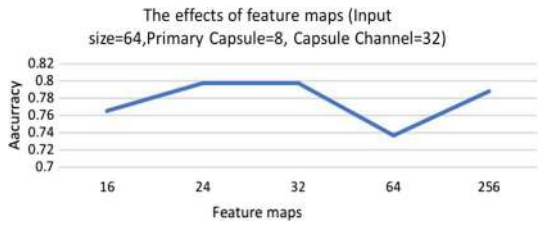

Fig. 3 The effects of feature maps in CapsNet

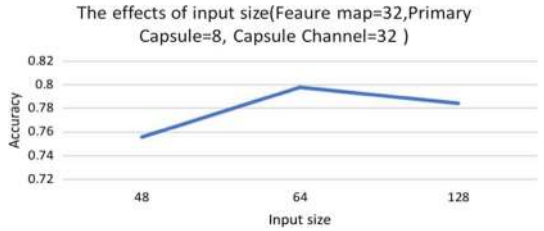

Fig. 4 The effects of input size in CapsNet

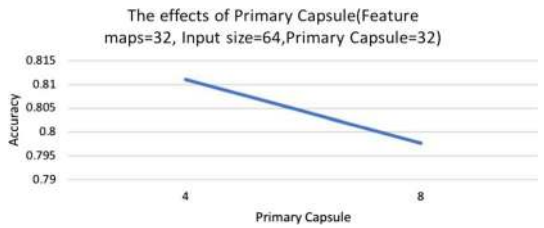

Fig. 5 The effects of primary capsule in CapsNet

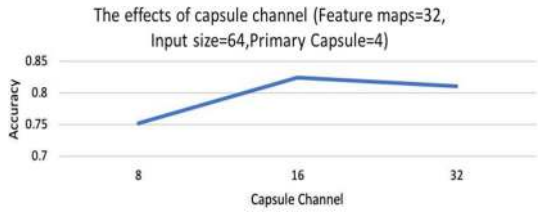

Fig. 6 The effects of capsule channel

3. The effects of primary capsule: A series of experiments can unveil the effects of feature maps through fixing feature maps size (32), input size (64), number of capsule channels (32) and varying the number of primary capsules. It can be seen in Fig. 5 that the model with primary capsule 4 got better accuracy than primary capsule 8.
4. The effects of capsule channel: A series of experiments can unveil the effects of feature maps through fixing feature maps size (32), input size (64), number of primary capsules (4) and varying the number of capsule channels. It can be seen in Fig. 6 that the model with capsule channel 16 got the best accuracy.

Table 16 Evaluation of dropout, batch normalization and learning rate for model2 with last unfrozen Convlayer

Model	Normal vs pneumonia			Bacteria vs virus		
	Accuracy	Specificity	Recall	Accuracy	Specificity	Recall
Baseline	0.776	0.809	0.776	0.643	0.64	0.585
VGG16 [34][a]	0.923	0.926	0.923	0.923	0.909	0.85
VGG16 [38][b]	0.938	0.944	0.938	0.915	0.917	0.879
Inception V3	0.869	0.854	0.869	0.851	0.86	0.779
CapsNet	0.824	0.846	0.824	0.862	0.875	0.785
Stateof-art [3][c]	0.928	0.901	0.932	0.907	0.909	0.886

[a] This result got from the version 2 of Kermany's dataset.

[b] This result got from the version 3 of Kermany's dataset that is a new released by authors to fix some error in version 2 dataset

[c] The state-of-art result got from the research of Kermany et al. [3], which is from a transfer learning based on InceptionV3

5. The best model: The model with a combination of feature maps size (32 or 24), input size (64), the number of the primary capsule (4) and the number of capsule channels (16) should get the best results. This can be verified by the results of test 11 and 13 in Table 11: they are the best of all the tests. This also agrees with the finding in transfer learning: The complexity of a model should match the scale of a dataset.

Horizontal comparison

To evaluate the performance of the models in this paper, Table 13 compares the best results of different models on the same pneumonia dataset. From Table 16, it can be seen that a neural network-based method is significantly better than the traditional method because it is a useful feature learner during the traditional method, just a feature-ORB. In version 2 of the dataset, the best model, VGG16 in this paper, got slightly lower accuracy and recall than the state-of-art result but obtained a higher specificity. On the latest dataset, the performance of VGG16 was generally higher. The VGG16 model released the last ConvLayer so that it would learn the specific features of the dataset. That should significantly help to improve performance very much. Kermany's work also retrains the ConvLayer of InceptionV3, but the model overfits too much to get an excellent test performance. The reason why our model does not overfit too much maybe because the VGG16 model is not as complicated as InceptionV3.

Finding in verifying on OCT dataset

From Table 12, it can be seen that the best model comes from test 5 instead of test 1. The new model adds complicated FC layers; therefore, the full complicity is better matched with the new dataset. The unfrozen two ConvLayers will make the system too complicated for the new dataset and, therefore, cannot find the local maxima. The best result is slightly lower than the start-of-art result of Kermany's work (96.6%). However, this experiment result also confirms our findings. The specific feature is most important to improve accuracy—the proper model complexity help to find the best result.

Conclusions and future work

Due to the importance of medical image classification and the particular challenge of the medical image-small dataset, this paper chose to study how to apply CNN-based classification to small chest X-ray dataset and evaluate their performance. From the experiments, the following finding presented. CNN-based transfer learning is the best method of all three methods. The capsule network is better than the ORB and SVM classifier. Generally speaking, CNN based methods are better than traditional methods because they can learn and select features automatically and effectively; The best results come from the transfer learning of VGG16 with one retrained ConvLayer, which is slightly higher than the start-of-art result. With the unfrozen ConvLayer, the specific feature can learn from the new dataset. Therefore, the specific feature is an essential factor to improve accuracy; The balance of a model's power of expression and overfitting is necessary. A too simple network usually cannot learn enough from the data, and therefore cannot get high accuracy. On the other hand, a very complex network is hard to train and tends to overfit quickly. As a result, accuracy is still low. Only a network model with proper size and other effective methods preventing overfit, such as proper dropout rate and proper data augmentation, can get the best results. However, because of the limited time, future research needs to be done: In transfer learning, training a fine-tuned deep neural network with unfrozen ConvLayers tends to overfit. What can effective methods be done to stabilize the training process? Other more powerful CNN model, such as ResNetv2 and ensemble of multiple CNN models have not been evaluated, but they could improve the results; Visualization needs to be added to improve the understanding and explanation of the results of the CNN-based system, because those are essential for the adoption of a CNN-based system in real clinical applications.

Abbreviations

CNN: convolutional neural network; VGG16: Visual Geometry Group from Oxford; ORB: oriented fast and rotated binary; SVM: support vector methods; DNN: deep neural networks; OCT: optical coherence tomography; GAN: generative adversarial networks; MNIST: Modified National Institute of Standards and Technology; k-NN: k-nearest neighbor; DWT: discrete wavelet transform; WFT: wavelet frame transform; WPT: wavelet packet transform; SIFT: scale invariant feature transform; CXR: chest X-ray; MRI: magnetic resonance imaging; SVHN: street view house numbers; CIAR10: Canadian Institute for Advanced Research 10; SSD: solid-state drive; GPU: graphics processing unit; VM: virtual machine; GAP: global average pooling; FC: fully connected.

Acknowledgements

We sincerely and gratefully acknowledge our organization Dr. Babasaheb Ambedkar Technological University for its help and support.

Authors' contributions

SSY and SMJ have contributed to the research with the order they appear. SMJ being the project principal investigator. Also, both authors discussed the final results as well as improved the final manuscript. Both authors read and approved the final manuscript.

Funding

This research did not receive any grant from any funding agencies in the public sectors.

Data availability statement

The dataset comes from the work of Kermnay et al. [34]. It contains two kinds of chest X-ray Images: NORMAL and PNEUMONIA, which are stored in two folders.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing of interests.

Author details

[1] Dr. Babasaheb Ambedkar Technological University, Raigad, Lonere, India. [2] Department of Information Technology, Dr. Babasaheb Ambedkar Technological University, Raigad, Lonere, India.

Received: 30 September 2019 Accepted: 3 December 2019
Published online: 17 December 2019

References

1. World Health Organization. Standardization of interpretation of chest radiographs for the diagnosis of pneumonia in children. Geneva: World Health Organization; 2001.
2. Center for Disease Control Prevention. Pneumonia can be prevented-vaccines can help. New York: Center for Disease Control Prevention; 2012.
3. Kermany DS, Goldbaum M, Cai W, Valentim CC, Liang H, Baxter SL, McKeown A, Yang G, Wu X, Yan F, et al. Identifying medical diagnoses and treatable diseases by image-based deep learning. Cell. 2018;172(5):1122–31.
4. Rawat W, Wang Z. Deep convolutional neural networks for image classification: a comprehensive review neural computation; 2017.
5. Jiménez-Sánchez A, Albarqouni S, Mateus D. Capsule networks against medical imaging data challenges. Intravascular imaging and computer assisted stenting and large-scale annotation of biomedical data and expert label synthesis. New York: Springer; 2018. p. 150–60.
6. Wang X, Peng Y, Lu L, Lu Z, Summers RM. Tienet: text-image embedding network for common thorax disease classification and reporting in chest X-rays. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2018. p. 9049–58.
7. Deng J, Dong W, Socher R, Li L-J, Li K, Fei-Fei L. Imagenet: a large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. 2009. p. 248–55.
8. Stanford Vision Lab, imagenet summary and statistics. http://www.image-net.org/about-stats. Accessed 30 Apr 2010.
9. Ding J, Chen B, Liu H, Huang M. Convolutional neural network with data augmentation for sar target recognition. IEEE Geosci Rem Sens Lett. 2016;13(3):364–8.
10. Frid-Adar M, Diamant I, Klang E, Amitai M, Goldberger J, Greenspan H. Gan-based synthetic medical image augmentation for increased cnn performance in liver lesion classification. Neurocomputing. 2018;321:321–31.
11. Vasconcelos CN, Vasconcelos BN. Increasing deep learning melanoma classification by classical and expert knowledge based image transforms. In: CoRR, vol. 1. 2017. arxiv:abs/1702.07025.
12. Zhou J, Li Z, Zhi W, Liang B, Moses D, Dawes L. Using convolutional neural networks and transfer learning for bone age classification. In: 2017 international conference on digital image computing: techniques and applications (DICTA). IEEE. 2017. p. 1–6.
13. Perez L, Wang J. The effectiveness of data augmentation in image classification using deep learning. 2017. arXiv preprint arXiv:1712.04621.
14. Shin H-C, Roth HR, Gao M, Lu L, Xu Z, Nogues I, Yao J, Mollura D, Summers RM. Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning. IEEE Trans Med Imaging. 2016;35(5):1285–98.
15. Goel R. Prediciting pneumonia with the help of transfer learning. 2018.
16. Sabour S, Frosst N, Hinton GE. Dynamic routing between capsules. In: Advances in neural information processing systems. 2017. p. 3856–66.
17. LeCun Y, Cortes C, Burges C. Mnist handwritten digit database. at&t labs. 2010.
18. Afshar P, Mohammadi A, Plataniotis KN. Brain tumor type classification via capsule networks. In: 2018 25th IEEE international conference on image processing (ICIP). IEEE. 2018. p. 3129–33.
19. Google, Google cloud gpus. 2018.https://cloud.google.com/gpu/.
20. Beutel J, Kundel H L, Van Metter R L, Fitzpatrick J M. Handbook of Medical Imaging: medical image processing and analysis, vol. 2. Bellingham: Spie Press; 2000.
21. Paredes R, Keysers D, Lehmann TM, Wein B, Ney H, Vidal E. Classification of medical images using local representations. In: Bildverarbeitung für die Medizin 2002. Berlin: Springer. 2002. p. 71–4.
22. Parveen N, Sathik MM. Detection of pneumonia in chest X-ray images. J X-ray Sci Technol. 2011;19(4):423–8.
23. Caicedo JC, Cruz A, Gonzalez FA. Histopathology image classification using bag of features and kernel functions. In: Conference on artificial intelligence in medicine in Europe. Berlin: Springer; 2009. p. 126–35.
24. Rublee E, Rabaud V, Konolige K, Bradski GR. Orb: an efficient alternative to sift or surf. Citeseer. 2011;11:2.
25. Mueen A, Baba S, Zainuddin R. Multilevel feature extraction and X-ray image classification. J Appl Sci. 2007;7(8):1224–9.
26. Yuan X, Yang Z, Zouridakis G, Mullani N. Svm-based texture classification and application to early melanoma detection. In: 2006 international conference of the IEEE engineering in medicine and biology society. IEEE. 2006. p. 4775–8.
27. Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, Huang Z, Karpathy A, Khosla A, Bernstein M, et al. Imagenet large scale visual recognition challenge. Int J Comput Vision. 2015;115(3):211–52.
28. Li Q, Cai W, Wang X, Zhou Y, Feng DD, Chen M. Medical image classification with convolutional neural network. In: 2014 13th international conference on control automation robotics & vision (ICARCV). IEEE; 2014. p. 844–8.
29. Wang X, Peng Y, Lu L, Lu Z, Bagheri M, Summers RM. Chestx-ray8: hospital-scale chest X-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2017. p. 2097–106.
30. Vianna VP. Study and development of a computer-aided diagnosis system for classification of chest X-ray images using convolutional neural networks pre-trained for imagenet and data augmentation. 2018. arXiv preprint arXiv:1806.00839.
31. Iesmantas T, Alzbutas R. Convolutional capsule network for classification of breast cancer histology images. In: International conference image analysis and recognition. Berlin: Springer. 2018. p. 853–60.
32. Beşer F, Kizrak MA, Bolat B, Yildirim T. Recognition of sign language using capsule networks. In: 2018 26th signal processing and communications applications conference (SIU). IEEE. 2018. p. 1–4.
33. Xi E, Bing S, Jin Y. Capsule network performance on complex data. 2017. arXiv preprint arXiv:1712.03480.

34. Kermany D, Goldbaum M. Labeled optical coherence tomography (oct) and chest X-ray images for classification. In: Mendeley data. 2018. p. 2.
35. James L. Setting up a Google cloud instance gpu for fast.ai for free. 2017.
36. N'úñez HJH. Python and opencv code for object classification using images. 2016.
37. Afshar P, Plataniotis KN, Mohammadi A. Capsule networks for brain tumor classification based on mri images and coarse tumor boundaries. In: ICASSP 2019–2019 IEEE international conference on acoustics, speech and signal processing(ICASSP). IEEE; 2019. p. 1368–72.
38. Kermany D, Goldbaum M. Labeled optical coherence tomography (oct) and chest X-ray images for classification. In: Mendeley Data. 2018. p. 3.

Publisher's Note