Improved Medical Image Classification Accuracy on Heterogeneous and Imbalanced Data using Multiple Streams Network

Mumtaz Ali¹
School of Computer Science,
Huazhong University of Science and Technology,
Wuhan & Department of Computer Systems Engineering,
Sukkur IBA University, Pakistan

Riaz Ali²
Department of Computer Science
Sukkur IBA University
Pakistan

Nazim Hussain³
Department of Mathematics
Sukkur IBA University
Pakistan

Abstract—Small and massively imbalanced datasets are long-standing problems on medical image classification. Traditionally, researchers use pre-trained models to solve these problems, however, pre-trained models typically have a huge number of trainable parameters. Small datasets are challenging for them to train a model adequately and imbalanced datasets easily lead to overfitting on the classes with more samples. Multiple-stream networks that learn a variety of features have recently gained popularity. Therefore, in this work, a quad-stream hybrid model called QuadSNet using conventional as well as separable convolutional neural networks is proposed to achieve better performance on small and imbalanced datasets without using any pre-trained model. The designed model extracts hybrid features and the fusion of such features makes the model more robust on heterogeneous data. Besides, a weighted margin loss is used to handle the problem of class imbalance. The QuadSNet is trained and tested on seven different classification datasets. To evaluate the advantages of QuadSNet on small and massively imbalanced data, it is compared with six state-of-the-art pre-trained models on three benchmark datasets based on Pneumonia, COVID-19, and Cancer classification. To assess the performance of QuadSNet on general classification datasets, it is compared with the best model on each of the remaining four datasets, which contain larger, balanced, grayscale, color or non-medical image data. The results show that QuadSNet handles the class imbalance and overfitting better than existing pre-trained models with much fewer parameters on small datasets. Meanwhile, QuadSNet has competitive performance in general datasets.

Keywords—Medical image classification; convolutional neural networks; class imbalance; small dataset; margin loss

I. INTRODUCTION

Typically, a huge amount of data is needed to train the neural networks for natural and medical image classification. However, along with the scarcity of sufficient samples, generally, the medical image datasets are massively imbalanced and they possess very limited positive cases. Thus, to obtain high-performance results by incorporating small and imbalanced datasets is a very perplexing task. In recent years, researchers have tried various algorithm-level and data-level methodologies to handle such challenges.

The algorithm level approaches have evolved since the reemergence of deep learning. Transfer learning [5], few-shot learning [19], zero-shot learning [6], Siamese networks [8], network ensembles [4] and most recent algorithms based on generative adversarial networks [2] have been applied to small and imbalanced datasets. The algorithm level approaches commonly rely on pre-trained models. Despite decent signs of progress, the pre-trained models orthodoxy use millions of parameters and complex architecture to achieve competitive results, which appear to be enormous for a small dataset to train a good model.

Along with the issue of a small dataset, the class imbalance is another challenge. Overmuch parameters of pre-trained models typically lead to overfitting if the classes are imbalanced. Weighting class labels is a very prevalent technique in handling class imbalance. The weighted class label approach produces better results when the model is trained with a suitable loss function and optimization method.

The data level approaches are mainly based on data augmentation, oversampling and undersampling. In medical image processing, conventional data augmentation techniques are sometimes problematic. For instance, a chest X-ray image traditionally depicts the heart at the lower left side, but when the image is augmented with a mirror effect, it will depict the heart at the right side of the chest. Similarly, effects of filling, zooming, or rotation may change the meaning of a medical image completely. Hence, efficiently designed deep learning models and algorithms may be a superior option for medical image processing.

However, despite good performance, data and algorithm level approaches tend to use very complex models with a large number of parameters. According to [3], recent hybrid models using multiple streams and feature fusion have produced competitive results for small and highly imbalanced datasets. In this work, the algorithm level approaches are followed and a quad-stream hybrid model, called QuadSNet, with separable as well as conventional convolutional neural networks is proposed to classify medical images on small and imbalanced datasets.

The performance of QuadSNet is evaluated together with the state-of-the-art pre-trained models, including DenseNet121 (DN) [12], InceptionV3 (IN) [25], MobileNet (MN) [11], ResNet50 (RN) [9], VGG16 (VG) [24], and Xception (XC) [7]. The proposed model QuadSNet is trained and tested on six medical image datasets and one famous MNIST handwritten digits dataset. QuadSNet is
intended to handle the problem of class imbalance with weighted margin loss on small datasets. Consequently, out of these seven datasets, the three smallest and most imbalanced medical datasets are used as benchmarks to compare the performance of QuadSNet with pre-trained models. For comprehensiveness, two of the three benchmark datasets are grayscale and one is colored. Other than the three benchmark datasets, the remaining four datasets are used to analyze the performance of QuadSNet on the larger, balanced, grayscale, color or non-medical image data.

The proposed technique could well handle the challenges of small datasets and class imbalance using a less complex model. Competitive results are obtained by using a fewer number of trainable parameters than most state-of-the-art pre-trained models. The feature fusion technique makes QuadSNet robust on a range of data. The training data size, type of images, number of classes and number of samples in each class are controlled with QuadSNet at an agreeable level.

The rest of the paper is organized as follows. In Section II, the proposed QuadSNet model and its architecture in detail is discussed. Section III presents a clear demonstration of the datasets, training procedure and experimental results. The best results in all the tables are in bold. Discussion and conclusion are in the end.

II. THE PROPOSED QUADSNET

The proposed QuadSNet model uses a quad-stream approach, as illustrated in Fig. 1. Out of the four streams, two streams use conventional convolutional neural networks [16], denoted as \( CS_1 \) and \( CS_2 \), and two streams use separable convolutional neural networks [7], denoted as \( SS_1 \) and \( SS_2 \), respectively. Each stream is based on blocks, whose detailed design is depicted in Fig. 4.

Each of the streams, \( CS_1, CS_2, SS_1 \) and \( SS_2 \), consists of four blocks. The design of the streams helps them to extract different features because of dissimilar kernel sizes and different forms of convolutional mode. Due to such a technique, the model becomes wider rather than deeper, and a wider approach helps to reduce the number of trainable parameters without losing accuracy. Features extracted from each stream are concatenated to make a fusion of the learned features. The model learns various features simultaneously. Therefore the training time is drastically reduced.

Separable convolutional neural networks are faster than conventional convolutional neural networks because of their depth-wise and point-wise feature extraction mechanism [7]. They perform fewer multiplications during operation than conventional convolutional neural networks. Therefore, both types of convolutional neural networks in are used in the model to make it faster and more robust. Model functions in a conventional manner as any convolutional neural network may operate. Thus, the quad-stream approach helps the model to learn features simultaneously on a single input image.

A. Feature Representation and Fusion

Formally, \( F_{CS_1}, F_{CS_2}, F_{SS_1} \) and \( F_{SS_2} \) are represented as the extracted features of the streams \( CS_1, CS_2, SS_1 \) and \( SS_2 \) respectively. The features \( F_{CS_1} \) and \( F_{CS_2} \) are generated through conventional convolutional neural networks, whereas \( F_{SS_1} \) and \( F_{SS_2} \) are the features generated through the streams of separable convolutional neural networks. The \( CS_1 \) and \( SS_1 \) streams use the kernel size of \( 3 \times 3 \) for conventional convolutional layers and separable convolutional layers respectively, and use \( 2 \times 2 \) for maxpooling layers. Similarly the \( CS_2 \) and \( SS_2 \) streams use the kernel size of \( 5 \times 5 \) for conventional convolutional layers and separable convolutional layers, respectively, and \( 2 \times 2 \) for maxpooling layers.

The features generated by \( CS_1 \) and \( CS_2 \) are concatenated into \( F_{cs} \). Likewise the features obtained through the streams \( SS_1 \) and \( SS_2 \) are concatenated into \( F_{ss} \). Finally the features \( F_{cs} \) and \( F_{ss} \) are concatenated into \( F_{total} \), as shown in Equation (1).

\[
F_{cs} = \begin{pmatrix} F_{CS_1} \\ F_{CS_2} \end{pmatrix}, F_{ss} = \begin{pmatrix} F_{SS_1} \\ F_{SS_2} \end{pmatrix}, F_{total} = \begin{pmatrix} F_{cs} \\ F_{ss} \end{pmatrix}
\] (1)

The feature fusion in \( F_{total} \) is eventually average-pooled to retain effective features and shrink the size of feature vectors.

B. Micromanagement of Over-fitting

The model handles over-fitting at the micro-level on each block of the model. Each block includes a dropout layer and a batch normalization layer. The dropout layer uses a fixed value of 0.3. The batch normalization layer depends upon two important parameters, momentum and epsilon. The momentum parameter is set to 0.99 and the \( \epsilon \) is set to 0.0001.

Other than the block-level dropout layers, there is a dropout layer after the average-pooling layer as well. Customarily the dropout layers using a value more than 0.5 in medical image classification are not considered in best practices. Therefore, it is purposefully kept at 0.3.

C. Margin Loss

The margin loss [28] with dynamic weights is adopted to handle the class imbalance and unnecessarily instant over-fitting of the model due to a smaller training set. Preliminary experiments showed that the margin loss performs better than the cross-entropy loss. Therefore, the margin loss is opted to be used, as represented in Equation (2).

\[
\Gamma _\kappa = T_\kappa max(0,m^+ - ||\nu _\kappa ||^2) + \lambda (1 - T_\kappa )max(0,||\nu _\kappa ||^2 - m^-)
\] (2)

In Equation (2) \( \kappa \) is the class present, \( m^+ = 0.9, m^- = 0.1 \) and \( \lambda = 0.5 \) is down-weighting. Here \( \nu _\kappa \) represents the feature vector with \( \kappa \) number of classes. The proposed model can be used not only for binary classification but also for multi-class classification. Therefore \( \kappa \) can be any finite discrete number. The values of \( m^+ \), \( m^- \) and \( \lambda \) dynamically handle the value of total loss \( \Gamma _\kappa \). Ultimately the loss is minimized gradually by controlling the imbalance among the \( \kappa \) classes.
III. Experiments

A. Datasets

The proposed QuadSNet model and six state-of-the-art pre-trained models are trained and tested on six different types of medical datasets and one non-medical image dataset. For convenience, the seven datasets are symbolically named, as Θ₁, Θ₂, ..., up to Θ₇ and the six pre-trained models as DN, IN, MN, RN, VG and XC, respectively.

1) Brief description of datasets: Each of the datasets in the list serves a unique purpose because of images’ types and classes’ degree of balance. Most datasets are imbalanced other than Θ₇, the MNIST dataset. Table I illustrates details of the datasets. The Θ₁, Θ₂, and Θ₃ are used as the benchmark datasets. Dataset Θ₁ consists of chest X-ray images in grayscale. Similarly, Θ₂ consists of CT-scan images in grayscale. Dataset Θ₃ is a colored dataset because of the dermoscopic images. As two of the benchmark datasets are grayscale and one is colored, the models have a fair chance of depicting efficacy. The QuadSNet and each of the six pre-trained models are trained on these datasets. The obtained results are then compared and analyzed comprehensively. Three benchmark datasets are described as follows:

1) Pneumonia (Θ₁): Θ₁ is a chest X-ray dataset for pneumonia classification [14]. This dataset is officially available at www.ijacsa.thesai.org.
as follows:

2) COVID-19 ($\Theta_2$): $\Theta_2$ is about COVID-19 (the data is from CT-scans) [29] and it is officially presented at https://covid-ct.grand-challenge.org.

3) Cancer ($\Theta_3$): Dataset $\Theta_3$ is RGB dataset based on dermatoscopic images for skin lesion classification [26]. The dataset is a random subset of the dataset present at https://challenge2019.isic-archive.com/data.html.

4) Hand Written Digit Classification MNIST ($\Theta_4$): This dataset is based handwritten digits [17] from 0 to 9. It is a multi-class classification data based on OCT (Optical Coherence Tomography) images of the retina of diabetic patients. This dataset contains four classes, called Choroidal Neovascularization (CNV), Diabetic Macular Edema (DME), DRUSEN, and Normal. This dataset is officially hosted at https://www.kaggle.com/paulimothymooney/kermany2018.

5) Malarial Cell Classification ($\Theta_5$): $\Theta_5$ is for Malarial cell classification. It is a balanced dataset [23]. This dataset is officially presented at https://www.kaggle.com/paulimothymooney/kermany2018.

6) Chest-xray-pneumonia/ (\Theta_6): The dataset is official and is available at https://www.kaggle.com/paulimothymooney/chest-xray-pneumonia/.

7) Breast Cancer (\Theta_7): Dataset $\Theta_7$ is about Breast Cancer (the data is from CT-scans) [29] and it is officially presented at https://covid-ct.grand-challenge.org.

Other than benchmark datasets, QuadSNet is also trained on $\Theta_4$, $\Theta_5$, $\Theta_6$, $\Theta_7$ to test its capabilities on a variety of image data. A brief description of the remaining datasets is presented as follows:

1) DR Classification on fundus photographs ($\Theta_4$): $\Theta_4$ is a relatively larger dataset than $\Theta_1$, $\Theta_2$ and $\Theta_3$; it is based on color fundus photographs. The data is officially presented at https://www.kaggle.com/c/dr-detect/data.

2) DR Classification on OCT images ($\Theta_2$): $\Theta_2$ is the largest dataset among all the datasets and it is massively imbalanced due to the very high number of normal samples and the smaller number of diseased samples [14]. It is a multi-class classification data based on OCT (Optical Coherence Tomography) images of the retina of diabetic patients. This dataset contains four classes, called Choroidal Neovascularization (CNV), Diabetic Macular Edema (DME), DRUSEN, and Normal. This dataset is officially hosted at https://www.kaggle.com/paulimothymooney/kermany2018.

3) Malarial Cell Classification ($\Theta_5$): $\Theta_5$ is for Malarial cell classification. It is a balanced dataset [23]. This dataset is officially presented at https://www.kaggle.com/paulimothymooney/kermany2018.

4) Hand Written Digit Classification MNIST ($\Theta_4$): This dataset is based handwritten digits [17] from 0 to 9. The dataset $\Theta_4$ is used to investigate the capabilities of QuadSNet on non-medical image datasets. Officially this dataset is present at http://yann.lecun.com/exdb/mnist/.

The main objective of using the seven different datasets is to analyze the diversity of the QuadSNet on a variety of image data. A few of the datasets are smaller and imbalanced except MNIST. The selected datasets are based on colored or grayscale images. Hence QuadSNet is tested accordingly.

B. Training and Testing

The QuadSNet and pre-trained models are trained and tested on a Windows 10 PC equipped with NVidia Gforce GTX 1060, having 16 GB of RAM, Intel Ci7 64 bit processor. All the simulations are performed on Keras with Tensorflow at the backend. The pre-trained models are individually trained on each of the three benchmark datasets.

Essential supplements and parameters, for instance, the learning rate $\gamma = 0.0001$, the optimizer, batch size $\beta = 32$, momentum $\mu = 0.009$, and input image size has been kept the same for all the models, including QuadSNet. Adam [15] optimizer has been used throughout the training. The models have been trained up to their maximum potential to nullify any unfairness.

Fig. 2 shows the training accuracy ($A$), the validation accuracy ($\Delta A$) of QuadSNet on all datasets, the final training loss ($\Gamma$) and validation loss ($\Delta\Gamma$) of QuadSNet on all datasets, the final training accuracy ($A$) and validation accuracy ($\Delta A$) Curves for $\Theta_1$, $\Theta_2$, and $\Theta_3$ of Different Models and (d) The Final Training Loss ($\Gamma$) and Validation Loss ($\Delta\Gamma$) Curves for $\Theta_1$, $\Theta_2$, and $\Theta_3$ of Different Models.

IV. DISCUSSION

The training accuracy $A$ and the validation accuracy $\Delta A$ of the pre-trained models and QuadSNet, as presented in Figure 2, exhibit the upshot of training with margin loss. As in this paper the weighted margin loss is used, it noticeably helps to control the effect of imbalance and scarceness of data with the help of QuadSNet. QuadSNet is trained from scratch. Therefore the micromanagement of over-fitting at the block level assists the model to handle it instantaneously.

The pre-trained models traditionally are trained on Ima-
geNet [10], which is mostly based on natural images. In the transfer learning approach, the intended dataset and type of images play a vital role in a model’s performance. If the results produced by pre-trained models are studied, datasets $\Theta_1$ and $\Theta_2$ were much easier because of precise edges and object-like structures. Whereas dataset $\Theta_3$ looks harder for the models to extract features due to the lack of natural objects like features.

The same conditions of pre-trained models apply to QuadSNet as well. Therefore it yields higher accuracy on datasets $\Theta_1$ and $\Theta_2$ than on dataset $\Theta_3$. The massive advantage of QuadSNet over pre-trained models in feature extraction is the fusion of multiple features obtained from various streams. Such features are rich due to different convolution techniques and different sizes of the kernels. The streams in QuadSNet are based on both separable and conventional convolutional neural networks; therefore, the features can be complementary with each other and contribute to the model’s accuracy on diverse data. Therefore, QuadSNet achieves a reasonable accuracy on heterogeneous image data.

A. Experimental Results

The systematic analysis of the experimental results of QuadSNet and the pre-trained models after the training reveals a massive difference between the number of trainable parameters. Table II shows the details about the number of trainable parameters for each of the models. QuadSNet and MobileNet have the fewest trainable parameters, which are about one-fifth of the average parameters.

Traditionally accuracy is not considered as a better performance metric for medical image classification on imbalanced data, therefore, sensitivity, specificity, and $F_1$ score are used.

All the datasets are distributed in three portions (Training, Validation, and Test). The main objective is to analyze the performance of QuadSNet in comparison with pre-trained models on imbalanced and smaller datasets. Therefore, the chosen three benchmark datasets, $\Theta_1$, $\Theta_2$, and $\Theta_3$, are the smallest and most imbalanced among the seven datasets.

The performances of pre-trained models and QuadSNet on $\Theta_1$, $\Theta_2$, and $\Theta_3$ are depicted in Table III. It can be concluded that QuadSNet achieves the best performance in all the benchmark datasets at almost all the metrics.

Except for the six state-of-the-art pre-trained models, QuadSNet is compared with some of the latest publications on each of the datasets which have better performance on the particular dataset. The results can be observed from Table IV. Each of the compared methods is designed for a particular disease. QuadSNet exhibits competitive accuracy, specificity and sensitivity in comparison with existing works. Most existing works listed in Table IV utilize pre-trained models having a huge number of parameters and use data augmentation techniques, whereas QuadSNet produces comparable results with a few trainable parameters and without any data augmentation. The dataset $\Theta_2$ is a non-medical image dataset. Therefore the results based on accuracy are compared only.

B. Limitations of the Study and Future Work

The proposed model has been thoroughly tested on the datasets described in Section III-A1. However, the proposed model has not been tested on 3-D medical image data, which may impact the performance. Similarly, the proposed model has not been tested in multi-input or multi-output settings where multiple tasks are performed simultaneously. In the future, the QuadSNet’s efficacy may be tested on multimodal or 3-D medical image data.

V. Conclusion

Compared with popular image datasets such as ImageNet for classification, medical datasets are usually small and class imbalance. This work introduces a new model called QuadSNet that can effectively handle over-fitting and class imbalance on small datasets without using transfer learning and data augmentation to handle such issues for image data. In general, QuadSNet outperforms the pre-trained models in terms of sensitivity, specificity, accuracy, and $F_1$ score on three medical benchmark datasets. Additionally, QuadSNet has competitive results compared to several state-of-the-art works focusing on a certain disease, demonstrating the effectiveness of the model. QuadSNet exhibits efficacy on a variety of datasets, including grayscale and color images. Due to the ability to handle diverse data, QuadSNet has the potential of
becoming a universal model for medical image classification on small datasets. This paper uses margin loss only, but QuadSNet can be trained and tested using other loss functions to analyze its effectiveness.

ACKNOWLEDGMENT

Thanks to Chao Li for organizing the paper and thanks to professor Kun He for guidance and support.

REFERENCES