Deep Neural Network using Hadoop for Cervical Cell Classification

Janani S[1] & Dr. D. Francis Xavier Christopher[2]

[1]*Research Scholar, Department of Computer Science,*
Rathnavel Subramaniam College of Arts and Science, Coimbatore, Tamil Nadu, India
[2]*Principal & Professor in Computer Science*
SRM Trichy Arts and Science College, Tiruchirapalli, Tamil Nadu, India

DOI: www.doi.org/10.34293/iejcsa.v2i10.10

Abstract - *Deep learning is a concept close to AI technique connecting neural networks. Several industries are poised to be revolutionized by deep learning.In this field, variety of data including images, sounds, and text are analysed through representing and abstracting information on multiple levels. The main objective of this work is to implement a ResNet like architecture in Hadoop for analysing cervical cancer big data to reduce processing time. Novel contribution includes implementing a distributed approach utilizing HDFS and MapReduce Frameworks in order to train ResNet-like neural networks in the field of cervical cell cancer classification.A big data platform called Hadoop is used to evaluate its implementation and performance. The proposed architecture applied to a preprocessed image of pap smear slide for classification and it yields good result for both 2-class (99.9% Specificity, 99.8% Accuracy, 99.7% h-mean and 99.1% Sensitivity) and 5-class classification problem (99.9% Specificity, 99.1% Accuracy, 99.6% h-mean and 99.1% Sensitivity).*

Keywords*: Neural Network, Distributed, Hadoop, Pap Smear, Cervical Cancer.*

INTRODUCTION

Advancement in IT resulted in great development in big data, thereby handling/processing this huge datasets using conventional techniques is tedious. In medical field, data being generated every day from various sources including pharmaceutical, clinical and laboratory data. Pathologist can provide solution by analysing the generated images in laboratory. But, it is a time consuming task to assess all the patterns in medical image by pathologist. Image classification task in image mining for various fields including healthcare, retail, food industry and agriculture becomes very easy by utilizing Convolutional Neural Networks (CNNs). Much proven architectures in previous work are well-established shows effective results for various tasks. In [1], an enhanced algorithm of TF-IDF is presented for reuters news retrieval. Results presented in [1] showed an improvement in map reduce based framework in news classification and clustering. In [2], to process large data in medicine, MapReduce technique is utilized. For smart cities [3] also, the model based on Hadoop framework can be used.

In [4], analysis of students and their behaviour as a model proposed using apriori algorithm. The model was executed in mapreduce framework.Similarly, [5] present a technique for sentiment analysis by utilizing a deep learning classifier and Hadoop framework. Proposed MeLoN (Multi-tenant deep Learning framework On yarN) [6] which runs applications of distributed deep learning on Hadoop. The main subject of [7] is the utilization of Hadoop by CNN. Two effective algorithms namely LSTM and Support Vector Machine (SVM) were used to handle traffic data to find out offending drivers. In [8], to handle large scale data, a system to detect traffic violation using mapreduce and deep learning is proposed. Effective solution made using this mapreduce based system. Krishnasamy et al., [9] utilized madreduce framework for the execution of Hybrid Density based Clustering and Classification (HDBCC) technique.Though several techniques were introduced, few issues including vanishing gradients and facilitating feature reuse are not addressed.

Deep learning becomes popular for enhancing image classification and recognition accuracy by solving more complex tasks. The problem of vanishing gradient and degradation have made training deep neural networks challenging [10]. ResNet (Residual Neural Network) is a leading architecture utilized for computer vision and image recognition tasks and solves gradient vanishing problem and degradation too.

The Hadoop framework is one of the most familiar technologies for processing big data. In a variety of areas, including behavior analysis, Hadoop and MapReduce techniques are used. Large-scale data can be processed in an efficient manner with a reduction in processing time using MapReduce model. In this paper, MapReduce based ResNetfor cervical cell classification is proposed.

The detail of remaining sections are as follows: In the second section, review of literature is presented. In 3rd section, proposed system is discussed and in 4th section, results obtained for the work is presented. In the last section(Section 5), conclusion is given.

METHODOLOGY

ResNet also known as "Residual Network," is a deep CNN technique. An important factor inResNet is its contribution in solving the gradient vanishing problem, which in turn limits the depth of NN. Skip Connections also called Residual Blocks are used in ResNet to connect block’s input to its output. ResNet acts as a basic structure for classifying images and also for several deep learning tasks.

ResNet on HDFS trains huge datasets permits to scale deep learning tasks to process huge data effectively. This can be used to train computationally demanding deep neural networks like ResNet. This complex process of training distributed across many nodes result in significant training time reduction. In HDFS, to train ResNet models, image datasets are stored and distributed effectively to multiple nodes. In case of node failure, lost data can be easily recovered by HDFS and ensures the continuation of process. Preprocessed data can also be stored in HDFS including augmented, resized, rotated etc., Performance, scalability, resource management and fault tolerance are few advantages of combining ResNet with HDFS while handling massive deep learning tasks.

In this study, ResNet-like architecture implemented with MapReduce framework to predict cervical cancer cells. The system designed is able to detect cervical cancer cells (2-class) and able to classify the classes(5-class). MapReduce is utilized with deep learning to process huge volume of data. In figure 1 and 2, the framework designed for the proposed system to solve two class and five class classification problem is shown. ResNet like architecture with MapReduce for cervical cancer prediction utilizing different layers illustrated in the figure. A preprocessed image is augmented and trained on deep neural networks based on layered training data. Dataset used is SipakMed (largest cervical cell dataset) from kaggle.

In this paper, (200,200,3) as an input size given while taking ResNet of simple version into account. It comprises of fourteen layers where first layer is given as 3×3 conv32. The proposed architecture uses 128 channels to minimize the cost of computation and time. Identical convolutional layers of three and block with two weight layers used in this structure. The proposed architecture is shown in table 1.Complex layers in ResNet extracts features in the input image (nucleus, cytoplasm, and background etc.,) for image categorization. In the slave node, cervical cell images are loaded and distributed. The image analysed from each slave not depend on other slave. Thus, slave nodes in map phase run the functions of cancer cell prediction. Similarly, reduce phase combine these layer to produce output.

Table 1: ResNet like Architecture

Architecture	Layer 1		Layer 2		Layer 3	
	Residual Block 1	Residual Block 2	Residual Block 1	Residual Block 2	Residual Block 1	Residual Block 2
Filter Size	3×3	3×3	3×3	3×3	3×3	3×3
Channels/Shape	32	32	64	64	128	128
Batch Normalization	True					

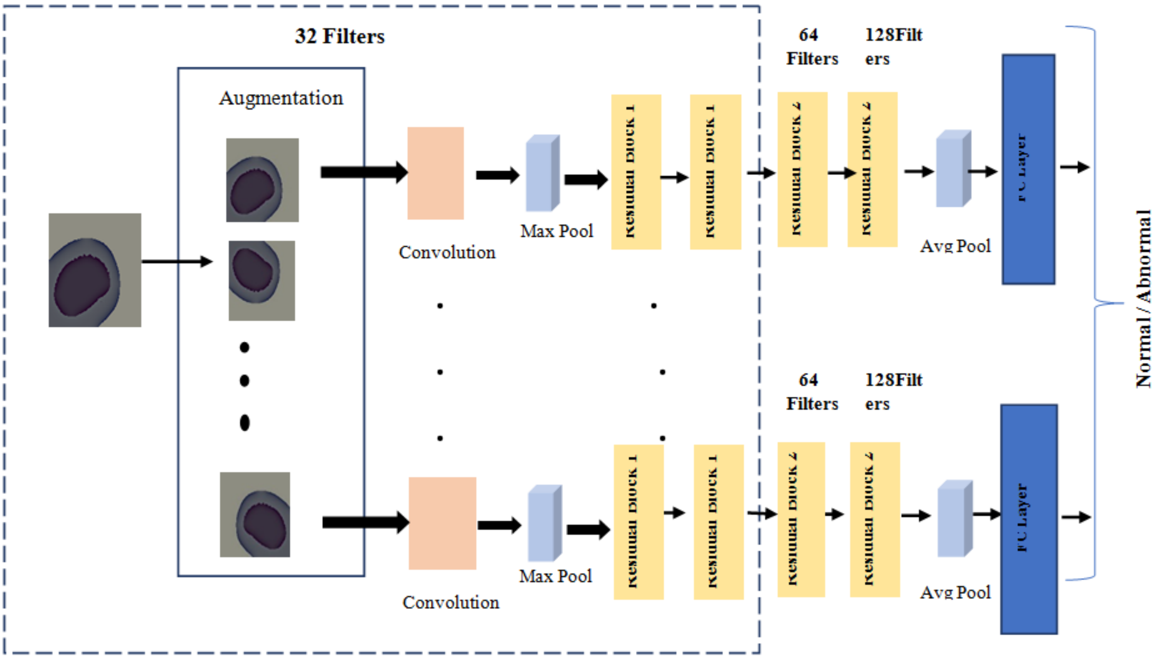

Figure 1 Two-Class Classification Problem using ResNet like Architecture on Distributed Environment

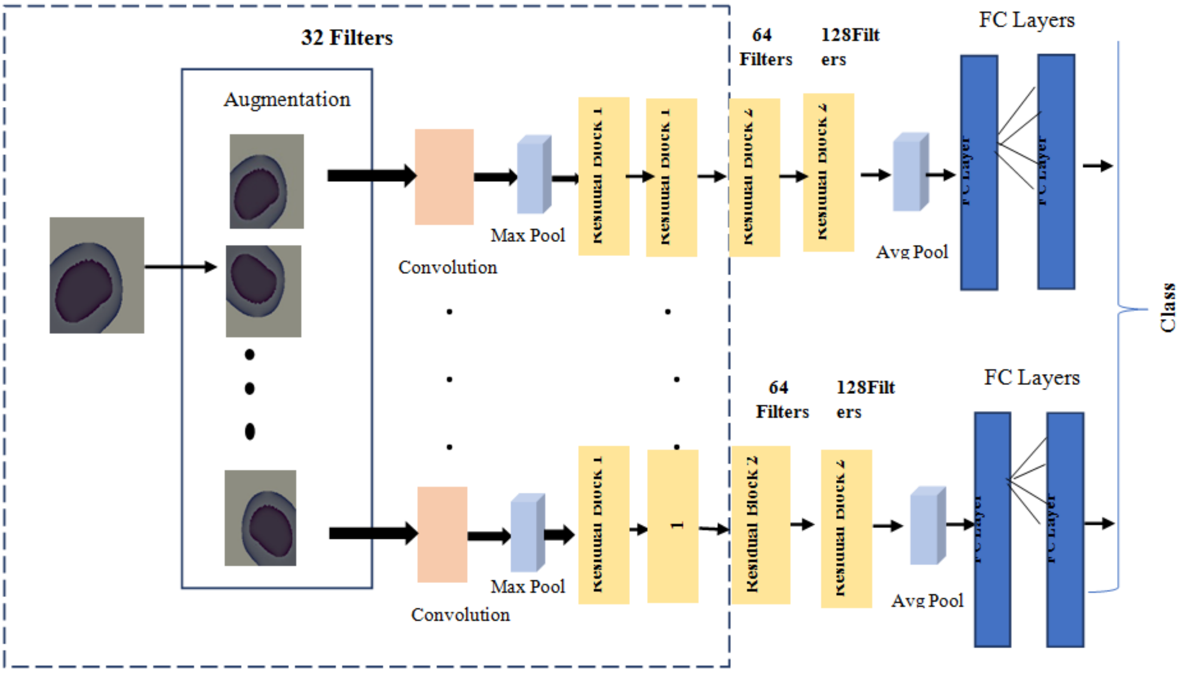

Figure 2 5-Class Classification Problem using ResNet like Architecture on Distributed Environment

Results and Discussions

In this sections, the results obtained by evaluating the proposed system is presented. Experiments were conducted using the pap smear slide dataset form kaggle.10,000 preprocessed images of size 200×200 across five classes were used. ResNet like architecture employed and changes done in distributed nodes, learning rates and batch size to evaluate the impact in the effectiveness of model. Python library Tensorflow, VMware environment for implementation were utilized to accomplish the proposed task. Results were carefully examined under various criteria including efficiency, scalability and performance etc., It is shown in table 2 that increasing the batch size impacts accuracy. Similarly, more nodes with smaller batch size provides good accuracy.

Table 2: Experimental Result under different Learning Rate and Batch Size

Experiment	Learning Rate	Batch Size	Distributed Node
1	0.001	64	2
2	0.01	128	4
3	0.001	32	8

ResNet based on distributed framework provides faster integration especially with 8 nodes. Number of nodes scaled up and shown improved performance. For evaluating scalability and performance, CPU time for executing camera images considered. With eight slave nodes in Hadoop cluster, the number of images for execution increased and CPU time reduced to almost 70%.

Efficiency: Images are loaded into HDFS for distributed processing. Input data split with MapReduce framework. In the map phase, a split of images taken by each mapper and previously trained ResNet model loaded. In ResNet'sConv layer, images are classified by feature extraction technique from the images. Class labels obtained as output. Similarly, in the reduce phase, final class labels are calculated for cervical cancer cells by combining the outputs of convolutional networks. Comparison made with 2-class classification problem in the existing work on different dataset since no attempt is made to use the SipakMed dataset to the best of our knowledge.

Result obtained using various metrics for evaluating proposed architecture for two class classification compared with existing work on different dataset is shown in table 3 and for five-class classification is shown in table 4. Proposed work applied for 2-class classification problem and yields accuracy of 99.8%, Specificity of 99.9%, Sensitivity of 99.1% h-mean of 99.7%. Similarly, accuracy of 99.1%, Specificity of 99.9%, Sensitivity of 99.1% and h-mean of 99.6% for 5-class classification problem. Generally, the described model based on MapReduce in Hadoop is effective for cervical cell classification.

Table 3: Comparison of Two-Class Classification Method on Different Dataset

Method	Specificity	Accuracy	Sensitivity	h-mean
Ensemble Learning [11]	83.59%	90.37%	96.33%	-
Deep Learning [11]	87.43%	91.63%	95.47%	-
VGG-like network [12]	99.9%	99.6%	98.8%	99.3%
Mean Shift+CAGA [13]	85.38%	82.17%	76.19%	-
Enhanced Fuzzy C-Means [14]	97.47%	98.88%	99.28%	
Our Method	99.9%	99.8%	99.1%	99.7%

Table 4: Five-Class Classification using Proposed Architecture

Specificity	Accuracy	Sensitivity	h-mean
99.9%	99.1%	99.1%	99.6%

ResNet's residual learning framework enables efficient training of very deep NN in distributed environments, overcoming issues like vanishing gradients and facilitating feature reuse. Its adaptability, superior generalization, and state-of-the-art performance make it a preferred choice, enhancing distributed systems' efficiency in handling diverse tasks and large datasets.

CONCLUSION

Huge data being generated in healthcare industry every day. Specifically, microscopic images play important roles in diagnosing diseases. Assessing huge volume of data by pathologist is time consuming and risk of giving inaccurate result. In this paper, preprocessed pap smear images are used for cervical cancer classification based on MapReduce framework to reduce the processing time. Proposed system utilizes ResNet like architecture to perform both two class and five class classification problem. Promising results shown using ResNet like architecture in distributed framework for image classification. The proposed architecture applied to a preprocessed image of pap smear slide for classification and it yields good result for both 2-class (99.9% Specificity, 99.8% Accuracy, 99.7% h-mean and 99.1% Sensitivity) and 5-class classification problem (99.9% Specificity, 99.1% Accuracy, 99.6% h-mean and 99.1% Sensitivity).The specific challenges faced in implementing ResNet in distributed systems, including communication optimization and hardware constrains, remain a knowledge gap. Additionally, comparative analysis with other architectures, real-world case studies, and updates on the latest ResNet variants in distributed learning are essential areas requiring further exploration.

REFERENCES

1. Chen, CH. 2017, 'Improved TFIDF in big news retrieval: An empirical study', Pattern Recognition Letters, vol. 93, pp. 113-122.
2. Kouanou, AT, et al. 2018, 'An optimal big data workflow for biomedical image analysis', Informatics in Medicine Unlocked, pp. 68-74.
3. Osman, AMS. 2018, 'A novel big data analytics framework for smart cities', Future Generation Computer Systems, vol. 91, pp. 620-633.
4. Cantabella, M, et al. 2019, 'Analysis of student behavior in learning management systems through a big data framework', Future Generation Computer Systems, vol. 90, pp. 262-272.
5. Khan, M, & Malviya, A. 2020, 'Big data approach for sentiment analysis of twitter data using hadoop framework and deep learning', International Conference on Emerging Trends in Information Technology and Engineering.
6. Kang, D, et al. 2021, 'MELON: Distributed deep learning meets the big data platform', IEEE International Conference on Autonomic Computing.
7. Asadianfam, S, et al. 2021, 'Hadoop deep neural network for offending drivers', Journal of Ambient Intelligence and Humanized Computing.
8. Asadianfam, S, et al. 2021, 'TVD-MRDL: Traffic violation detection system using MapReduce-based deep learning for large-scale data', Multimedia Tools and Applications.
9. Krishnaswamy, R, et al. 2023, 'Metaheuristic based clustering with deep learning model for big data classification', Computer Systems Science and Engineering, vol. 44, no. 1, pp. 391-406.

10. Reddy, ASB & Juliet, DS. 2019, 'Transfer learning with ResNet-50 for malaria cell-image classification', International Conference on Communication and Signal Processing.

11. Kuko, M & Pourhomayoun, M. 2020, 'Single and clustered cervical cell classification with ensemble and deep learning methods', Information System Frontiers.

12. Kurnianingsih, et al. 2019, 'Segmentation and classification of cervical cells using deep learning', IEEE Access.

13. Wang, P, et al. 2019, 'Automatic cell nuclei segmentation and classification of cervical Pap smear images', Biomedical Signal Processing and Control, vol. 48, pp. 93-103.

14. William, W, et al. 2019, 'Cervical cancer classification from Pap-smears using an enhanced fuzzy C-means algorithm', Informatics in Medicine Unlocked, vol. 14, pp. 23-33.