Histopathological Image Classification: Defying Deep Architectures on Complex Data

Suvidha Tripathi(✉) and Satish Singh

Indian Institute of Information Technology Allahabad,
Devghat, Jhalwa, Allahabad 211015, U. P., India
suvitri24@gmail.com, sk.singh@iiita.ac.in

Abstract. Automatic analysis of medical images is a challenging research which requires both the skill of a pathologist and computer vision knowledge to develop efficient systems. In this work, we have taken up the task of classifying different types of cell nuclei in histopathological Colon Cancer Images. We aim to show the relevance and effect of a complex histopathological dataset on the performance of current deep learning architectures. We have experimented with pre-trained (on ImageNet) AlexNet, VGG16, and VGG19 architectures and applied transfer learning approach to train these architectures. On the basis of the results obtained on the Histopathological image dataset, while using fine tuned AlexNet, VGG16, and VGG19 architectures; the suitability of using pure architectures is somehow questionable and these state of the art algorithms straightaway cannot be used for the sophisticated classification of very complex cancer tissue dataset. Comparative evaluation of the above state of the art methods have been done and the possibility of devising hybrid deep architectures is investigated thereof.

Keywords: Transfer learning · Traditional features ·
Neural networks · Deep learning · Histopathology · Nuclei classification

1 Introduction

Traditional Technique of extracting features from a set of images is becoming obsolete and as a result, works claiming to use it are being taken lightly unless the authors make a thorough comparison with deep architectures. However, this is also true that no one has actually been able to come up with any such handcrafted or object level feature descriptors that has outperformed deep neural networks. One thing that gets remarkably get overlooked is the type of diversity of dataset that is being used to produce high accuracies in this domain. Histopathological or medical data, in general, is one such dataset that has been challenging researchers around the world due to their heterogeneity and uncertain nature of patterns. In this work, we have taken some of the very famous

K. C. Santosh and R. S. Hegadi (Eds.): RTIP2R 2018, CCIS 1036, pp. 361–370, 2019.
https://doi.org/10.1007/978-981-13-9184-2_33

deep learning methods to do the classification task using transfer learning technique and highlight the fact that even the most efficient networks may fail to improve the classification accuracy when the dataset involved is highly complex in nature. Starting from AlexNet [1], VGG16, and VGG19 [2], three of the most used architectures due to their low computational requirements and high performance, are used in this experimental study.

The histopathological dataset used here is a set of different types of nuclei found in colon cancer i.e., fibroblasts, inflammatory, lymphocytes, neutrophils, eosinophils, etc. The presence of each type of these cell nuclei indicates the nature of cancer [4]. This information is crucial for the pathologists to diagnose the severity and type of cancer. The factors that influence the decision of diagnosis are size, structure, density, chromatin texture and intensity (depends on the staining dye) of the nuclei present in the affected tissues [5]. These factors change with the type of nuclei and hence, pathologists need to know beforehand the type to make a conscious decision. The properties of this dataset are crucial since such type of datasets changes with the type of staining technique used to stain the nuclei and stroma of the tissues. Hence, different staining technique gives different color and texture features. So to summarize, color and texture features play a very important role in classifying different structures, even more than the shape and size features.

With deep learning architectures we can very easily extract the features after each layer but, lack of interpretability restricts us to know the actual type of features extracted in the process. Therefore, it is difficult to firmly establish whether the quality of features deep learning network is extracting will give good classification performance.

In past years, much work has been done on histopathological images for nuclei classification using handcrafted feature descriptors such as morphological, texture, shape and color features. Shape representation through DTW-Radon based descriptors by authors in [3] and [6] estabilised a rotation and scaling invariant lossless transform for detecting various shape properties in several numeral, character and symbol datasets. Their methods could be used to detect shape features in nuclei datasets for classification purposes but, since the dataset is very huge and contains large number of samples, their method would incur huge computational cost. Liu et al. in [7] tested the various types of features that can be extracted from images and through feature selection methods they found out the most relevant features for cell nuclei classification. However they did not mention the kind of dataset they used for extracting features, therefore it is hard to say if their findings are universal for all types of datasets. Authors in [8] studied various nuclei classification methods on different types of cancer kinds such as prostate, breast, renal clear cell and renal papillary cell cancer. They showed that the classification methods gave different accuracies on each one of these datasets. Their results proved that there cannot be one definitive method that would give better results across all types of cancer. They also used deep learning methods like LeNet, EncoderNet, Color-EncoderNet on their datasets but only Color-EncoderNet was able to give best results among the

10 methods they tested on 3 out of 4 datasets. These studies hint that even using the deep learning framework does not guarantee good results in the case of complex histopathological images. Other effective methods such as [10] and [11] to detect candidate region of interests and extracting features for further processing of biomedical images use hand-crafted feature descriptors and local variations within the images. These methods are effective for small datasets that have greater inter and less intra class separability. But, in our case, where the dataset is complex and has less inter class separability, relying only on hand-crafted features is not a feasible approach. To prove this, we have tested few state of the art algorithms on hand-crafted feature descriptors and compared the obtained results with deep learning methods. Deep learning method used in [12] would have been an initial approach for classification but their method use grayscale images and also the dataset used is not from biomedical domain. So, the complexity and feature relation is highly deviated from our intensity and color centric RGB histopathological dataset.

2 Experiments

We have taken some of the very recent deep architectures and trained our dataset on them to find out their performance.

2.1 Dataset

Image dataset from which nuclei points are extracted as patches is taken from [9]. The dataset came with annotated nuclei and their location in the data. We prepared our own data points using the method in [14]. From each image in the dataset as shown in Fig. 1, the nuclei present in this image were annotated by pathologists. Annotated nuclei center pixel coordinates were recorded for each of the images along with their corresponding labels. Using this recorded information about all the nuclei, total 22444 nuclei samples of height and width 27 around the center pixel coordinate, with RGB color channels, were collected in a folder. 22444 nuclei were segregated into four classes viz. Epithelial nuclei, Inflammatory nuclei, Fibroblast nuclei and miscellaneous other types as the fourth category. The number of samples in each class affect the final results by a great margin. In our dataset, class 1 i.e. epithelial class has total 7,722 nuclei, class 2 (inflammatory nuclei) has 5,712 samples, class 3 (fibroblast nuclei) has 6,971 class points and the miscellaneous category has mixed type data of total 2,039 sample points. These raw nuclei images were then divided into train and test set. 70% of the samples from each class were taken as input for training and the rest 30% of the samples were used in testing. However, The input size of each image in our dataset had to be resized to $224 \times 224 \times 3$ since, this is the size that the AlexNet, VGG16 and VGG19 architectures take as input. Figure 2 shows the sample nuclei dataset.

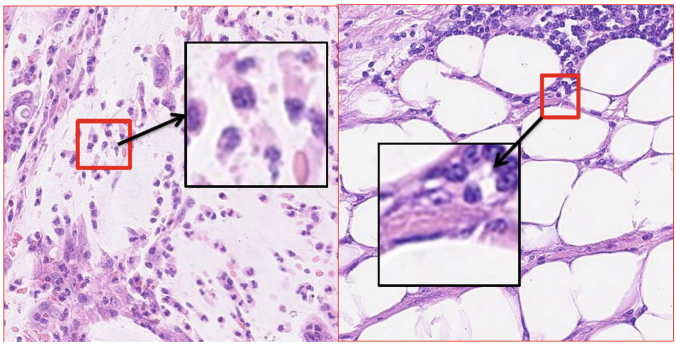


Fig. 1. 500 × 500 H&E stained histology image samples of colorectal adenocarcinomas

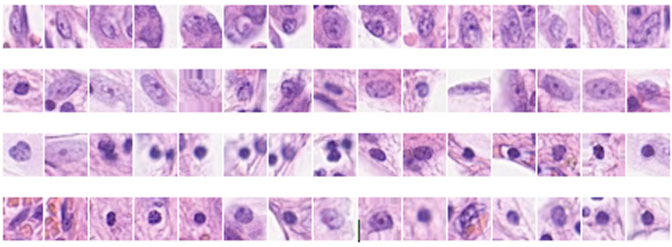


Fig. 2. Example of nuclei dataset. Row 1: epithelial nuclei, Row 2: inflammatory nuclei, Row 3: fibroblasts, Row 4: miscellaneous

2.2 AlexNet

AlexNet by Krizhevsky et al. [1] is the very first architecture inspired from LeCun et al. [15] which gained popularity after 2012 ImageNet challenge. It has 5 convolutional layers followed by 3 fully connected layers. We divided the dataset into 7:3 ratio for training and testing. Initially, we kept the learning rate incremental, starting from 0.01 and increased it up to 0.00001 i.e. 1e−05. we observed the minibatch accuracy very low and the overall accuracy on pretrained AlexNet was observed 0. So, incremental learning rate did not work with our dataset. Hence, we kept the learning rate constant.

2.3 VGG16

We investigated the effect of increasing depth of convolutional layers by testing the performance of VGG16 [2] on our dataset accuracy. The number of parameters increases with the depth and hence the computation requirements. We trained our dataset using the pre-trained model because learned features are often transferable to different data and then it also takes less training time as compared to the experiment where the model is trained from the scratch [13].

Training any deep learning architecture from scratch is not feasible for both accuracy and time performance since the network has to learn again the trivial features like edges and lines which becomes a redundant task if the accuracy does not improve as the training progress. Using the concept of transfer learning helps propagate the generic features through the model. Only the features specific to the dataset are learned through model training.

2.4 VGG19

VGG19 [2] has more depth than VGG16 i.e. 19 convolutional layers and hence, improved performance. Working on this theory we trained VGG19 on our dataset and made few observations included in Sect. 3.

Apart from transfer learning, random changes in batch size and number of epochs were performed to select the optimal hyperparameters. We selected the batch size of 300 and trained the architectures for 100 number of epochs.

3 Results and Discussions

We have evaluated our classifier performance using Precision (or Positive Predictive Value PPV), Recall (or True Positive Rate TPR), F1 score, Accuracy and time taken by three architectures. Accuracy and time comparison among three architectures are shown in Table 1. It is observed that with deep architectures large batches can be parallelized across many machines, reducing training time significantly. Also, large batch size reduces the number of parameter updates required to train a model which in turn results in reduced model training times. Therefore, we kept the batch size high. To establish our design choice of a large batch size we did random batch size changes, starting from 64. We noticed no change in accuracy but, the time required to train each batch increased by 100%. Earlier, the time for each epoch, in case of 300 batch size, was around 20 min, which increased to 42 min when the batch size was reduced to just 64 and number of epochs to 30. This happens when lower batch size takes a number of iterations to do the weight update due to more number of computations. So, It was more feasible to train our dataset with a larger batch size considering the time efficiency. The recent article by authors in [21] have studied the effect of increasing batch sizes on ImageNet and CIFAR10 datasets using recent state of the art deep learning algorithms like ResNet and Inception-ResNet-V2. They confirmed that the large batch size reduces the training times significantly and are better than decaying learning rate when the effect on accuracy is not significant. Figures 3a, b and c are the ROC curves of three networks. Each figure has four curves representing four classes of nuclei i.e., Epithelial, Fibroblast, Inflammatory and miscellaneous. We have also compared our deep learning architecture performance with handcrafted descriptors we used in [14] to measure their retrieval performance on our dataset. Comparison Table 5 clearly outlines the fact the handcrafted descriptors are clearly no match to deep learning algorithms since

Table 1. Accuracy and time comparison

Architectures	Accuracy (%)	Time (secs)
Alexnet	72.68	2564.69
VGG16	73.89	9331.79
VGG19	73.54	59976.00

there is a huge difference in classification metrics. While the same descriptors performed better in retrieving CT, MRI, and ultrasound images such as in [16–18], they performed very poorly on our dataset when we used the same feature subset for classification. It is important to note that the feature descriptors specially designed to retrieve medical images in [16–18] performed even poorer than the ones that were designed for retrieving colored images [19,20]. So, it establishes the fact the color information is an important feature in case of histopathological images. Handcrafted features that work on grayscale images will not give optimum performance in such datasets.

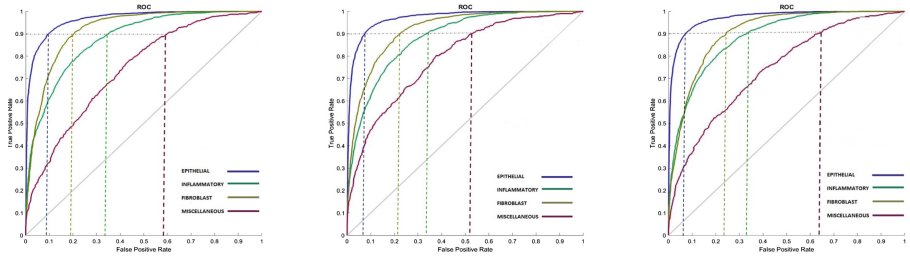
Table 2. Confusion matrix of AlexNet

Output class	Target class				
	1	2	3	4	Precision
1	2017	261	145	93	80.9%
2	152	1104	280	138	65.9%
3	52	198	1524	233	75.9%
4	6	151	142	148	33.1%
Recall	90.9%	64.4%	72.9%	24.2%	72.5%

We made following observations from the results we obtained.

1. ROC curves are shown in Fig. 3 shows the performance of each architecture AlexNet (Fig. 3a), VGG16 (Fig. 3b), and VGG19 (Fig. 3c). To compare the differences among these curves we took True Positive Rate (TPR) value at 90% in all three curves and noticed the corresponding False Positive Rate (FPR) with respect to each class. FPR value should be minimized with respect to each class. In case of class 1 (Epithelial) minimum, FPR is given by VGG19 and maximum FPR is by AlexNet whereas for class 4 (Miscellaneous) minimum and maximum FPR is given by VGG16 and VGG19 respectively. For Inflammatory nuclei category, FPR is almost similar in all three methods and in the class of Fibroblast nuclei, VGG16 gives the minimum FPR and VGG19 outputs maximum FPR. After analyzing the three ROC curves, we inferred that there is no unique pattern to declare the best classifier for all 4 classes. They show different patterns with respect to each class. This difference in

patterns may become a problem when determining the best classifier among the three. However, due to an imbalance in the data samples, it is expected that the fourth class which has the least number of samples will perform the worst. This gives the clue to the best classifier question, which is, the classification method that performs the best with minority class should be the best classifier. Here VGG16 has the minimum FPR with minority class. Hence, VGG16 is the best classifier among the three. This is also reflected in the classification metrics Table 5.



(a) AlexNet (b) VGG16 (c) VGG19

Fig. 3. ROC curves of AlexNet, VGG16 and VGG19

2. if we compare our results with ImageNet dataset accuracies on these networks, that is AlexNet has top 1 accuracy of 56.1% [1] and top 5 accuracies of 80%, VGG16 has top 1 and top 5 accuracies of 70.6% and 89.9%, and for VGG19 it is 68% and 85.5% respectively [2], we see that there is a significant improvement of atleast 12% in top 1 accuracy and 6% increase when comparing top 5 accuracies of AlexNet and VGG19.
3. Hence, by observation of accuracy changes among datasets, we can very certainly say that our dataset was indeed difficult to classify for these architectures.
4. We also made observations among class wise accuracy, and uniformly we noticed from confusion matrices that class 1 i.e. Epithelial nuclei scored the best with highest percentage of 84.9% in case of VGG19 (Table 2). Class 2 (Inflammatory nuclei) second with the highest percentage of 65.9% in AlexNet (Table 3), class 3 (Fibroblasts) third with highest 76.6% in VGG16 (Table 4) and miscellaneous nuclei in class 4 scored the least accuracy among all three architectures with best value of 42.7% in VGG16 (Table 4).
5. This variation in accuracies reflect on the structure of the nuclei in the database. Miscellaneous nuclei contained all other small groups of nuclei found in colon cancer, hence this class did not have any particular pattern in majority. Therefore, the classifier could not make the best decision for this class.
6. We observed from Table 5 that despite VGG19 having the deepest network, did not perform better than VGG16. But, it is however not a very significant improvement. VGG16 is only 1% more sensitive (recall) than VGG19

Table 3. Confusion matrix VGG19

Output class	Target class				
	1	2	3	4	Precision
1	2119	191	117	68	84.9%
2	145	1224	411	128	64.2%
3	42	191	1488	295	73.8%
4	11	108	75	121	38.4%
Recall	91.5%	71.4%	71.2%	19.8%	73.5%

Table 4. Confusion matrix of VGG16

Output class	Target class				
	1	2	3	4	Precision
1	2132	208	145	91	82.8%
2	138	1214	386	146	64.4%
3	40	181	1489	234	76.6%
4	7	111	71	141	42.7%
Recall	92.0%	70.8%	71.2%	23.0%	73.9%

Table 5. Comparison between methods through performance parameters

Method	Precision	Recall	F1-score	Accuracy
LBDP	38.80%	31.45%	34.74%	40.7%
LCOD	48.27%	38.00%	42.52%	46.20%
LWP	38.06%	31.05%	34.20%	39.30%
LDEP	37.30%	30.15%	33.34%	39.10%
RSHD	49.30%	36.65%	42.04%	44.10%
AlexNet	63.95%	63.10%	63.52%	72.50%
VGG16	**66.62%**	**64.25%**	**65.41%**	**73.90%**
VGG19	65.34%	63.47%	64.39%	73.50%

(Table 5). Also, when we look at the time took by VGG16 and VGG19 from Table 1 for training, VGG19 took 6 times more time than VGG16. Hence, if we have to choose between VGG16 and VGG19, VGG16 becomes the better choice both in terms of accuracy and time.

7. From the comparison of the handcrafted and deep learning architectures in Table 5, it is trivial to deduce that deep architectures performed better than handcrafted descriptors used in this study.

4 Conclusion

Through this experimental work, our objective was to establish that, the state of the art deep learning networks perform better than handcrafted features but may not produce great results for all kinds of datasets such as the Histopathological data whereas, AlexNet, VGG16, and VGG19 produces classification accuracy better in ImageNet dataset as mentioned in point 2 of Sect. 3. Histopathological data is highly complex and incomprehensible to the non-experts. Without the consultation of domain expertise of the experienced pathologists, one can never be sure of the nature of the objects present in the images. Hence, proper classification of the images is a complex task even for humans for such datasets. Deep learning algorithms do not address the dataset heterogeneity problem and their performance in different domains of data. Therefore, with our work, we have tried to reflect on the fact that otherwise widely used deep learning algorithms used for classifying histopathological data are not the best feasible methodology alone. Hence, only handcrafted or only deep learning architectures are not enough for classifying complex histopathological data. Their combination shall be exploited to achieve the better performance.

Acknowledgment. This research was carried out in Indian Institute of Information Technology, Allahabad and supported by Ministry of Human Resource and Development, Government of India. We are also grateful to the NVIDIA corporation for supporting our research in this area. Currently, we are using a donated TITANX(PASCAL) GPU with 3584 CUDA cores to train models for this research work.

References

1. Alex, K., Ilya, S., Geoffrey, H.: Image net classification with deep convolutional neural network. In: NIPS (2012)
2. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
3. Santosh, K.C., Lamiroy, B., Wendling, L.: DTW-radon-based shape descriptor for pattern recognition. Int. J. Pattern Recogn. Artif. Intell. **27**(03), 1350008 (2013)
4. https://www.cancerresearchuk.org/what-is-cancer/how-cancer-starts/types-of-cancer
5. Zink, D., Fischer, A.H., Nickerson, J.A.: Nuclear structure in cancer cells. Nat. Rev. Cancer **4**, 677–687 (2004). https://doi.org/10.1038/nrc1430
6. Santosh, K.C., Lamiroy, Bart, Wendling, Laurent: DTW for matching radon features: a pattern recognition and retrieval method. In: Blanc-Talon, Jacques, Kleihorst, Richard, Philips, Wilfried, Popescu, Dan, Scheunders, Paul (eds.) ACIVS 2011. LNCS, vol. 6915, pp. 249–260. Springer, Heidelberg (2011). https://doi.org/10.1007/978-3-642-23687-7_23. Inria-00617287
7. Liu, S., Mundra, P.A., Rajapakse, J.C.: Features for cells and nuclei classification. In: 2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society, Boston, MA, pp. 6601–6604 (2011). https://doi.org/10.1109/IEMBS.2011.6091628

8. Singh, M., Zeng, Z., Kalaw, E.M., Giron, D.M., Chong, K.-T., Lee, H.K.: A study of nuclei classification methods in histopathological images. In: Chen, Y.W., Tanaka, S., Howlett, R., Jain, L.C. (eds.) InMed 2017. SIST, vol. 71, pp. 78–88. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-59397-5_9
9. Sirinukunwattana, K., Raza, S.E.A., Tsang, Y.W., Snead, D.R.J., Cree, I.A., Rajpoot, N.M.: Locality sensitive deep learning for detectionand classification of nuclei in routine colon cancer histology images. IEEE Trans. Med. Imaging **35**(5), 1196–1206 (2016)
10. Santosh, K.C., Wendling, L., Antani, S., Thoma, G.R.: Overlaid arrow detection for labeling regions of interest in biomedical images. IEEE Intell. Syst. **31**(3), 66–75 (2016). https://doi.org/10.1109/MIS.2016.24
11. Ravi, M., Hegadi, R.S.: Detection of Glomerulosclerosis in diabetic nephropathy using contour-based segmentation. In: International Conference on Advanced Computing Technologies and Applications ICACTA (2015)
12. Ukil, S., Ghosh, S., Obaidullah, S.M., Santosh, K.C., Roy, K., Das, N.: Deep learning for word-level handwritten Indic script identification, arXiv preprint arXiv:1801.01627
13. Yosinski, J., Clune, J., Bengio, Y., Lipson, H.: How transferable are features in deep neural networks? In: Advances in Neural Information Processing Systems 27 (NIPS 2014). NIPS Foundation (2014)
14. Tripathi, S., Mishra, S., Singh, S.K.: Routine colon cancer detection using local image descriptors. In: IEEE Region 10 Conference (TENCON), Singapore 2016, pp. 2062–2065 (2016). https://doi.org/10.1109/TEN-CON.2016.7848388
15. LeCun, Y., et al.: Backpropagation applied to handwritten zip code recognition. Neural Comput. **1**(4), 541–551 (1989)
16. Dubey, S.R., Singh, S.K., Singh, R.K.: Local diagonal extrema pattern: a new and efficient feature descriptor for CT image retrieval. IEEE Signal Process. Lett. **22**(9), 1215–1219 (2015)
17. Dubey, S.R., Singh, S.K., Singh, R.K.: Local bit-plane decoded pattern: a novel feature descriptor for biomedical image retrieval. IEEE J. Biomed. Health Inform. **20**(4), 1139–1147 (2015)
18. Dubey, S.R., Singh, S.K., Singh, R.K.: Local wavelet pattern: a new feature descriptor for image retrieval in medical CT databases. IEEE Trans. Image Process. **24**(12), 5892–5903 (2015)
19. Dubey, S.R., Singh, S.K., Singh, R.K.: Rotation and scale invariant hybrid image descriptor and retrieval. Comput. Electr. Eng. **46**, 288–302 (2015)
20. Dubey, S.R., Singh, S.K., Singh, R.K.: Local neighbourhood-based robust colour occurrence descriptor for colour image retrieval. IET Image Process. **9**(7), 578–586 (2015)
21. Smith, S.L., Kindermans, P.J., Ying, C., Le, Q.V.: Don't decay the learning rate, increase the batch size. In: ICLR 2018, arXiv preprint arXiv:1711.00489