Multi-Part Knowledge Distillation for the Efficient Classification of Colorectal Cancer Histology Images

Shankey Garg
Department of Computer Science and Engineering
National Institute of Technology Raipur,
Raipur, Chhattisgarh, India
sgarg.phd2019.cse@nitrr.ac.in

Pradeep Singh
Department of Computer Science and Engineering
National Institute of Technology Raipur,
Raipur, Chhattisgarh, India
psingh.cs@nitrr.ac.in

***Abstract* - Colorectal cancer is the most common type of cancer after breast cancer in women and third in men after lungs and prostrate cancer. The disease rank third in incidence and second in terms of mortality, hence early diagnosis is necessary for the correct line of treatment. Knowledge distillation based models boost the performance of small neural network and are performing efficiently for various image classification based tasks. In this work, a novel knowledge distillation based technique is developed to efficiently classify colorectal cancer histology images. Unlike traditional distillation, out method performs distillation in parts. Instead of supervising the student with a converged knowledge of teacher, the proposed method is fetching the teacher's knowledge at regular intervals and providing these knowledge to the student model during student training process. Through this multi-part distillation technique student can effectively learn the intermediate representational knowledge rather than the abstract knowledge of the teacher and hence boost the overall performance of the model. The the proposed model has achived 92.10% accuracy.**

Keywords- Classification, Knowledge Distillation, Colorectal Cancer, Histology Images

I. INTRODUCTION

The Global Cancer statistics 2020 estimates the cancer incidence and mortality to be around 1.9 million new cases and around 10.0 Million deaths[1].Colorectal cancer stands third in incidence while it ranks second in terms of mortality[2]. Colorectal cancer is commonly developed from dysplastic adenomatous polyps consisting of multiple steps like inactivation of various genes resulting in tumor suppression, repairing of DNA, and simultaneously activating oncogenes. In 99% of the cases, colorectal cancer occurs in people aged more than 40. Along with the age, family history is also the most common risk factor for colorectal cancer. Heavy alcoholism, smoking, and the consumption of red/processed meat are some of the additional risk factors[1][3]. Routine checkups may be an important step for reducing mortality rates of cancer patients. For the colorectal cancer diagnosis, visual assessment of the colorectal cancer whole slide images is considered the most prominent method in clinical practice. But the manual evaluation of these slides is a time taking process in the clinical routine[4][5].

In literature, Deep learning-based methods are mostly used by the researchers of the field to analyze the histopathological images[6][7]. But, only a handful of literature covers various aspects of colorectal cancer classification and detection using deep learning techniques. Further, it is required that the models developed for the classification of colorectal cancer could be deployed in resource-constrained devices for real-time applications. To the best of our knowledge, no studies suggest the development of these kinds of models for classifying colorectal cancer.Making the models lighter by reducing the number of parameters that could easily be deployed in the resource-constrained device is a key challenge[8]. It is necessary to reduce the number of parameters from the trained model as it may take more time to process the input during inference. Model compression and acceleration techniques have already attracted researchers inthe last couple of years and many techniques are gradually developed in this area, some of them are, low-rank factorization, knowledge distillation, network pruning, and quantization[9][10]. Among these techniques, Knowledge distillation (KD) efficiently optimizes the model by developing a light and efficient model. In this technique, the lightweight and small student networks mimic the behavior and the soft outputs of a powerful teacher network[11][9][12].KD is a primary mechanism, that enables humans to learn new complex concepts using small training sets having same or different categories[13]. But the key challenge in implementing knowledge distillation lies in how to transfer the knowledge from a cumbersome teacher model to a small and light student model. The key contributions of the paper is summarized in the following two point:

- In this work, distilled knowledge of teacher is transfered to the student model in parts at

978-1-6654-9291-1/22/$31.00 ©2022 IEEE

regular intervals for the efficient learning of the student.

- Proposed Multi-Part KD technique is used for the efficient classification of colorectal cancer histology images and the results are compared with the vanilla KD technique.

II. Related Work

A. Disease Classification

Deep learning has become the centre of attraction for the researchers working in the field of image classification, segmentation etc. It is advantageous for various problems in the field of health imformatics and, bioinformatics [14][15][16]. In the field of medical imaging, histopathological image plays an important role. Muhammad Talo [17], proposed an efficient transfer learning based approach to classify histopathology images using different pre-trained models. Singh et al. [18], addressed the imbalanced issue of data and proposed a transfer learning based framework for the classification of breast cancer histopathology images. Further, a light weight deep learning based model is developed by Garg and Singh [19] to address the imbalanced issue of breast cancer histology images and outstanding results were obtained using a light model containing less parameters and small model size. Ghosh et al.[20] proposed an ensemble deep neural network based method to efficiently classify colorectal cancer histology images using two publicly available datasets.

B. Knowledge Distillation

Knowledge Distillation was first proposed by Hinton et al. [11] in the year 2015. KD is a technique that helps in the training of a smaller student model under the guidance of a cumbersome teacher model. In this technique, the knowledge of teacher is given to the student by mininmizing the difference between the logits generated by the teacher model and the student model[12][13]. Ruffy and chahal[21] verified various KD techniques for classification tasks and implemented some recent techniques to examines the effectiveness of KD in classification based tasks. They observed that improvement is gained in these technique when tuned appropriately using traditional distillation strategy in combination with data augmentations. Fu et.al.[22] proposed an effective distillation technique by randomly performing swapping in the teacher blocks to replace the student blocks in each step and named the method as interactive knowledge distillation.A novel method for the compression of deep learning model is developed by combining weight pruning and knowledge distillation for the head-pose estimation regression network and image classification network by Aghli and Ribeiro[23]. According to Tan et. al.[24], if a teacher has weak ability to learn the knowledge of true data, then the student could not learn knowledge from its teacher. Hence they proposed an interclass correlation regularization for the training of teacher that could capture more explicit correlation among classes. Jin et. al.[25], proposed an efficient method for boosting the performance of student network by developing an easy to hard sequence of learning target. By this method, the student network could outperform compared with other knowledge transfer methods. This method was implemented for image classification and face recognition tasks. From the comprehensive literature survey it could be concluded that task of transfereing knowledge from teacher to student model plays a key role in the process of knowledge distillation. In this paper, our objective is to efficiently transfer the knowledge of teacher to student model for better learning.

III. Background and Notation

A. Knowledge Distillation

The aim of knowledge distillation is to train a lightweight student model under the supervision of an efficient and cumbersome teacher model. This knowledge transfer is carried out by imitating the soft labels to learn the dark knowledge observed by a teacher[26]. This dark knowledge is transferred in the form of logits z an output of the last fully connected layer of deep network. For any random i^{th} class, the logits are represented as z_i, then the probability p_i that input belongs to the ith class is calculated by the softmax function :

$$p_i = \frac{\exp(z_i)}{\sum_j \exp(z_j)} \tag{1}$$

Hence, these predictions of soft targets from teacher model contains dark knowledge and it is used as a supervisor for transferring knowledge from teacher to student. Each soft target is smoothened by introducing a temperature factor τ in the above equation:

$$p_i = \frac{\exp\left(\frac{z_i}{\tau}\right)}{\sum_j \exp\left(\frac{z_j}{\tau}\right)} \tag{2}$$

Here, the higher temperature produces a softer probability distribution over classes. When, $\tau \to \infty$, all classes share same probability and for $\tau \to 0$, these soft targets become hard targets, i.e. one-hot labels.

Given a teacher network T_n having parameters $\emptyset_T$ and a student network S_n with parameters $\emptyset_S$. The output predictions of the teacher and student are represented as

T_n= softmax (z_t) and S_n=softmax(z_s) respectively, where z_t and z_s are the logits of teacher and students. The objective is to let the student mimic the behavior of the teacher by minimizing the knowledge distillation loss that is represented as follows:

$$\mathcal{L}(\emptyset_S) = \mathcal{H}(y, S_n) + \lambda\mathcal{H}\,(T_n^{\tau}, S_n^{\tau}) \qquad (3)$$

In the above equation, $\mathcal{H}(\cdot,\cdot)$ denotes cross entropy and λ is hyperparameter used for balancing the two terms. τ in the T_n^{τ}, S_n^{τ} denotes a relaxation parameter known as temperature.

B. Proposed Multi-Part Knowledge Distillation

Multi-Part knowledge distillation is implemented using the intermediate training states by setting checkpoints at regular intervals. In the proposed work, checkpoints are set after every 5 epochs and a total of 10 intervals are considered. These checkpoint are represented as $C = C_1, C_2, \ldots, C_{50}$. With respect to the above mentioned checkpoints, output is represented as $\emptyset_t(x; S_{C_1}), \emptyset_t(x; S_{C_2}), \ldots, \emptyset_t(x; S_{C_{50}})$. Step-by-step training of student is performed by mimicking the checkpoints from teacher training until reaching the last checkpoint. The aim at any random i[th] checkpoint is denoted by:

$$\mathcal{L}(S_n, S_{C_i}) = \mathcal{H}\big(y, \emptyset_S(x; S_n)\big) + \lambda\mathcal{H}\,\Big(\emptyset_S(x; S_n), \emptyset_t\big(x; S_{C_i}\big)\Big) \qquad (4)$$

Learning target for any random i[th] step become $\emptyset_t(x; S_{C_i})$ of i[th] check point. Here $i \in \{1,2,\ldots,50\}$ and S_n is optimized by learning checkpoints sequentially.

C. Check-Point Selection

It is possible to produce any number of checkpoints from teacher network during its training. But, it is important to find an optimal strategy for this checkpoint selection. For this work, a simple and straight forward strategy is adopted that supervises the student network at equal epoch interval i.e. one checkpoint after every 5 epochs.

D. Backbone Network

It is important to select an efficient and feasible model as a backbone network for the student teacher architecture. In this work, variants of MixNet is adopted for teacher and student networks.This architecture focuses on the effect of kernel size for depthwise convolution and MixConv, capable of mixing multiple kernels in a single op that are simple drop-in replacements of depthwise convolution. The fundamental difference between vanilla depthwise convolution and MixConv is that vanilla depthwise convolution uses a single kernel for all the channels. In contrast, MixConv partitions channels into groups and applies different kernel sizes to each group. Novel versions of MixNets were developed using the neural architecture search method like MixNet-S, MixNet-M, MixNet-L, and MixNet-XL[27]. In this work, pre-trainedMixNet-L is used as teacher networks while MixNet-S is used for the student network.

IV. METHODOLOGY

Overall framework of the proposed multi-part knowledge distillation is shown in figure 1. The hypothesis for performing this knowledge distillation at regular intervals is to make the student network learn about the representational knowledge from the teacher rather than the abstract knowledge that is generated by the vanilla KD task.In the proposed framework, teacher network is given the input images that is trained for 50 epochs. Objective of proposed work is to divide the teacher training process into several checkpoints and collect the knowledge at each checkpoint. The knowledge collected from these checkpoints during teacher's training is used in the student network at the same intervals as used in the teacher's training process. Finally the performance of student network is evaluated after the final epoch.For this work, the entire process is divided into 10 equal intervals after every 5 epochs. Hence the teacher is storing knowledge for student after every 5[th] epoch and these knowledge is fed to the student model during the student training phase after every 5[th] epochs. Using the representational knowledge of the teacher, the performance of student network is evaluated at the end of 50[th] epoch. In this work, pre-trained MixNet-L is used for teacher network and the lighter version of teacher i.e. untrained MixNet-S is used for student training.

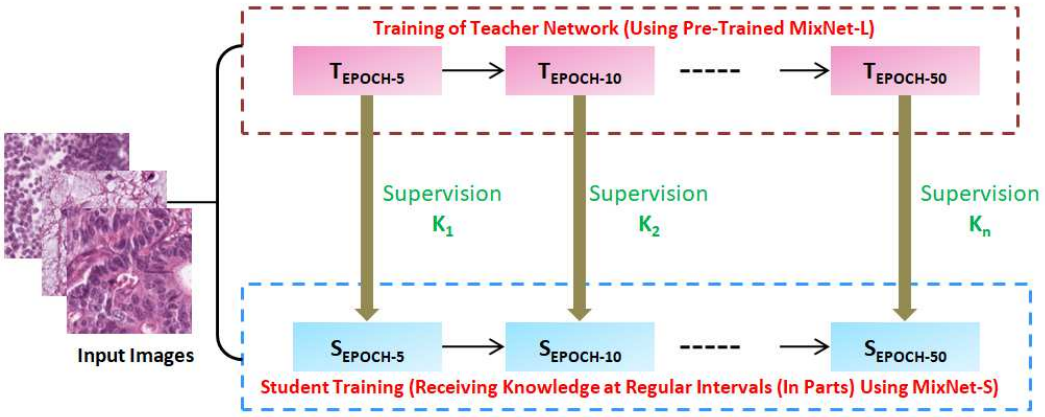

Figure 1. Proposed Multi-Part Knowledge Distillation Framework (K represents the knowledge generated from the checkpoints)

V. EXPERIMENTS

A. Dataset Used

University Medical Center Mannheim (Heidelberg University, Mannheim Germany) collected ten H & E (haematoxylin and eosin) stained colorectal tissue slides from the pathology lab[4]. These slides were digitized, and then contiguous tissue areas were annotated manually, creating 625 non-overlapping tissue tiles having dimensions 150px * 150px. Eight different tissues were identified for analysis in this collection. Table 1 gives the details of the tissue types and count of the images present in each class, and figure 2 depicts the image samples belonging to each class. For the evaluation, we have divided the dataset into train (80%), validation (20% of training data) and test (20%) sets. All the images were resized from 150 x 150 to 112 x 112 and divided into a batch size of 128. Essential normalization, resizing, and random horizontal flip are performed for all the training images. The test set images are resized and normalized only. Finally, the processed images are fed to the model for evaluation of the results.

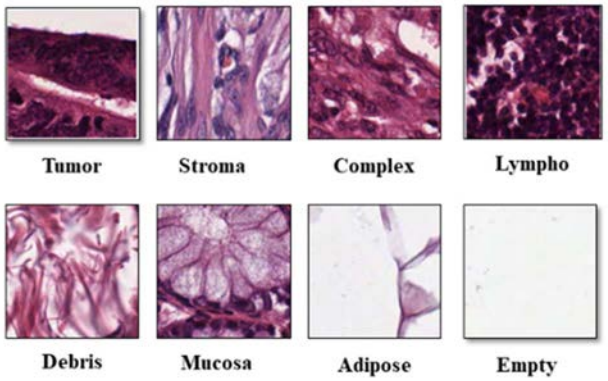

Figure 2. Image Samples Belonging to Each Class of the Colorectal Histology Images Dataset

B. Experimental Settings

Backbone network for teacher's training is MixNet-L and MixNet-S, a more compact version of theMixnet-L is used in the student network. For both the teacher and student networks, initial learning rate is set to 0.001, with a batch size of 64. ADAM optimizer is used with step size of 5 and gamma value is set to 0.5. Further, there are two most important parameters that is to be decided while performing KD related operations, these are temperature and distill weights. For this work the temperature value is set to 5 and the value of distill weight is set to 0.5. Teacher network uses cross entropy loss while the student network uses the combination of cross entropy loss with temperature, distill weights and KL Divergence loss along with the outputs of teacher and studentsAll experiments were conducted using PyTorch deep learning library in Google colab with the NVIDIA-SMI 495.46, CUDA Version: 11.2, Tesla K80, 12.69 GB RAM and 78.19 GB hard disk.

Table I. Details of the Dataset with Image Count of Each Class

Class Name	No. of Images
Tumor Epithelium	625
Simple Stroma	625
Complex Stroma	625
Immune Cells	625
Debris	625
Normal Mucosal Glands	625
Adipose Tissue	625
Background	625
Total	5000

VI. RESULT ANALYSIS

Comprehensive result evaluation is perfromed for the proposed work using the publicly available colorectal cancer histology images. This dataset is evaluated using the proposed KD technique and the conventional vanilla KD technique to check the effectiveness of the proposed KD method. Results obtained from the proposed KD and vanilla KD is depicted in table II. From table II it can be observed that the proposed Multi-Part KD techniqueboosts the performance approximately

1.64%as compared to vanilla KD. Further, it is important to discuss the model size and parameter requirements of the teacher and student network since we are utilizing a network compression technique (KD) and the amount of compression done in the student network and the analysis of the effect of compression on the model's performance is also a very important. These parameters are also discussed in table II.

Table II Performance Evaluation of the Proposed Multi-Part KD

Method	Network	Accuracy	Number of Parameters	Model Size
Teacher	MixNet-L	92.58	5.8M	23.218MB
Student	MixNet-S	88.02	2.6M	10.44MB
Vanilla KD	MixNet-S	90.46	2.6M	10.44MB
Proposed Multi-Part KD	**MixNet-S**	**92.10**	**2.6M**	**10.44MB**

VII. DISCUSSION

The Proposed work shows the effectiveness of knowledge distillation based method for the efficient classification of colorectal cancer histology images. In recent years, extensive research is being conducted in the field of transfer learning and deep learning but the area of light weight deep learning models and other network compression techniques are yet to be explored for the tasks related to cancer classification. This work is an attempt to uitilize knowledge distillation technique for classifying histology images. The evaluation of proposed technique is done by comparing the performance with vanilla KD for colorectal histology image classification. From the table II, it observed that proposed work outperforms vanilla KD in terms of accuracy while there is only a minor loss of accuracy when comparing the proposed technique with cumbersome teacher network. Further, it is seen that the student model is approximately two times lighter than the teacher model in terms of model parameters and model size.

VIII. CONCLUSION

In this work a simple and effective KD method is proposed to boost the performance of small student network by performing multi-part distillation technique. The proposed technique is performing better than the simple vanilla KD and the performance of small student network is almost near to that of teacher network. In future, we would like to explore some methods for finding the best checkpoints for the proposed method. Also, it is required to test the proposed model with different benchmark dataset to check the generalizability and adaptability of the work.

References

[1] F. Sung, Hyuna and Ferlay, Jacques and Siegel, Rebecca L and Laversanne, Mathieu and Soerjomataram, Isabelle and Jemal, Ahmedin and Bray, "Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries," *CA. Cancer J. Clin.*, vol. 71, no. 3, pp. 209–249, 2021.

[2] A. Siegel, Rebecca L and Miller, Kimberly D and Goding Sauer, Ann and Fedewa, Stacey A and Butterly, Lynn F and Anderson, Joseph C and Cercek, Andrea and Smith, Robert A and Jemal, "Colorectal cancer statistics, 2020," *CA. Cancer J. Clin.*, vol. 70, no. 3, pp. 145–164, 2020.

[3] A. B. Ballinger and C. Anggiansah, "Colorectal cancer," *Br. Med. J.*, vol. 335, no. 7622, pp. 715–718, 2007.

[4] F. G. Kather, Jakob Nikolas and Weis, Cleo-Aron and Bianconi, Francesco and Melchers, Susanne M and Schad, Lothar R and Gaiser, Timo and Marx, Alexander and Zollner, "Multi-class texture analysis in colorectal cancer histology," *Sci. Rep.*, vol. 6, pp. 1–11, 2016.

[5] L. D. Tamang and B. W. Kim, "Deep learning approaches to colorectal cancer diagnosis: A review," *Appl. Sci.*, vol. 11, no. 22, 2021.

[6] V. Rachapudi and G. Lavanya Devi, "Improved convolutional neural network based histopathological image classification," *Evol. Intell.*, vol. 14, no. 3, pp. 1337–1343, 2021.

[7] S. and others Wang, Kuan-Song and Yu, Gang and Xu, Chao and Meng, Xiang-He and Zhou, Jianhua and Zheng, Changli and Deng, Zhenghao and Shang, Li and Liu, Ruijie and Su, "Accurate diagnosis of colorectal cancer based on histopathology images using artificial intelligence," *BMC Med.*, vol. 19, no. 1, pp. 1–12, 2021.

[8] Y. Zhou, S. Chen, Y. Wang, and W. Huan, "Review of research on lightweight convolutional neural networks," *Proc. 2020 IEEE 5th Inf. Technol. Mechatronics Eng. Conf. ITOEC 2020*, no. Itoec, pp. 1713–1720, 2020.

[9] T. Choudhary, V. Mishra, A. Goswami, and J. Sarangapani, *A comprehensive survey on model compression and acceleration*, vol. 53, no. 7. Springer Netherlands, 2020.

[10] B. L. Deng, G. Li, S. Han, L. Shi, and Y. Xie, "Model Compression and Hardware Acceleration for Neural Networks: A Comprehensive Survey," *Proc. IEEE*, vol. 108, no. 4, pp. 485–532, 2020.

[11] G. Hinton, O. Vinyals, and J. Dean, "Distilling the Knowledge in a Neural Network," *arXiv Prepr. arXiv1503.02531*, vol. 2, pp. 1–9, 2015.

[12] J. Gou, B. Yu, S. J. Maybank, and D. Tao,

"Knowledge Distillation: A Survey," *arXiv*, 2020.
[13] L. Wang and K. J. Yoon, "Knowledge Distillation and Student-Teacher Learning for Visual Intelligence: A Review and New Outlooks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 14, no. 8, pp. 1–38, 2021.
[14] and G.-Z. Y. Daniele Rav`ı, Charence Wong, Fani Deligianni, Melissa Berthelot, Javier Andreu-Perez, Benny Lo, "Deep Learning for Health Informatics," *IEEE J. Biomed. Heal. Informatics*, vol. 21, no. 1, pp. 4–21, 2017.
[15] H. P. Chan, R. K. Samala, L. M. Hadjiiski, and C. Zhou, "Deep Learning in Medical Image Analysis," *Adv. Exp. Med. Biol.*, vol. 1213, pp. 3–21, 2020.
[16] D. S. W. Ting, Y. Liu, P. Burlina, X. Xu, N. M. Bressler, and T. Y. Wong, "AI for medical imaging goes deep," *Nat. Med.*, vol. 24, no. 5, pp. 539–540, 2018.
[17] M. Talo, "Automated classification of histopathology images using transfer learning," *Artif. Intell. Med.*, vol. 101, p. 101743, 2019.
[18] A. Seemendra, R. Singh, and S. Singh, "Imbalanced Breast Cancer Classification Using Transfer Learning," *IEEE/ACM Trans. Comput. Biol. Bioinforma.*, vol. 694, no. 1, pp. 425–436, 2021.
[19] S. Garg and P. Singh, "Transfer Learning Based Lightweight Ensemble Model for Imbalanced Breast Cancer Classification," *IEEE/ACM Trans. Comput. Biol. Bioinforma.*, vol. 5963, no. c, pp. 1–1, 2022.
[20] S. Ghosh, A. Bandyopadhyay, S. Sahay, R. Ghosh, I. Kundu, and K. C. Santosh, "Colorectal Histology Tumor Detection Using Ensemble Deep Neural Network," *Eng. Appl. Artif. Intell.*, vol. 100, no. February, p. 104202, 2021.
[21] F. Ruffy and K. Chahal, "The State Of Knowledge Distillation For Classification Tasks," *arXiv*, pp. 1–8, 2019.
[22] S. Fu, Z. Li, Z. Liu, and X. Yang, "Interactive Knowledge Distillation for image classification," *Neurocomputing*, vol. 449, pp. 411–421, 2021.
[23] N. Aghli and E. Ribeiro, "Combining weight pruning and knowledge distillation for CNN compression," *IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Work.*, pp. 3185–3192, 2021.
[24] C. Tan, J. Liu, and X. Zhang, "Improving knowledge distillation via an expressive teacher," *Knowledge-Based Syst.*, vol. 218, p. 106837, 2021.
[25] X. Jin, Xiao and Peng, Baoyun and Wu, Yichao and Liu, Yu and Liu, Jiaheng and Liang, Ding and Yan, Junjie and Hu, "Knowledge distillation via route constrained optimization," *Proc. IEEE Int. Conf. Comput. Vis.*, vol. 2019-Octob, pp. 1345–1354, 2019.
[26] K. Zhang, C. Zhanga, S. Li, D. Zeng, and S. Ge, "Student Network Learning via Evolutionary Knowledge Distillation," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 8215, no. c, pp. 1–13, 2021.
[27] M. Tan and Q. V. Le, "MixConv: Mixed depthwise convolutional kernels," *30th Br. Mach. Vis. Conf. 2019, BMVC 2019*, 2020.