Comparison Study of Deep Learning Models for Colorectal Lesions Classification

Berk Cetinsaya
University of Central Florida
4000 Central Florida Blvd
Orlando, FL 32816
+1(407) 823-2000
berkcetinsaya@gmail.com

James Dials
Florida Polytechnic University
4700 Research Way
Lakeland, FL 33805
+1(863) 583-9050
jdials3259@floridapoly.edu

Doga Demirel
Florida Polytechnic University
4700 Research Way
Lakeland, FL 33805
+1(863) 583-9050
ddemirel@floridapoly.edu

Tansel Halic
University of Central Arkansas
201 Donaghey Ave
Conway, AR 72035
+1(501) 450-5000
tanselh@uca.edu

Suvranu De
Renesselaer Polytechnic Institute
110 8th St
Troy, NY 12180
+1(518) 276-6000
des@rpi.edu

Mark Gromski
Indiana University School of Medicine
340 W 10th St #6200
Indianapolis, IN 46202
+1(317) 274-8157
mgromski@iu.edu

Douglas Rex
Indiana University School of Medicine
340 W 10th St #6200
Indianapolis, IN 46202
+1(317) 274-8157
drex@iu.edu

ABSTRACT

In this paper, we performed a comparison study between GoogLeNet, AlexNet, and InceptionV3 deep learning models to recognize and classify colorectal cancer tumors. Colorectal tumors are one of the very common cancer types and early detection could result in a significantly higher survival rate of 95% as opposed to 12%. In this work, we aim to investigate the deep learning models to automatically detect the tumor types from polyp images. We, therefore, used actual images taken from the colorectal surgery or colonoscopy using Narrow-band imaging (NBI). The images are classified based on NBI International Colorectal Endoscopic (NICE) classification. We used NICE 1 and NICE 2 types with a total of 2604 images in the size of 64x64. Our results show that the InceptionV3 model has the most accurate results by average 92.39% where AlexNet is 88.19% and GoogLeNet is 85.73%.

CCS Concepts

•Computing methodologies~Machine learning~Machine learning approaches~Neural networks

Keywords

Image Recognition; Deep Learning; Convolutional Neural Network

ICISDM 2020, May 15–17, 2020, Hawaii, HI, USA

ACM ISBN 978-1-4503-7765-2/20/05…$15.00
https://doi.org/10.1145/3404663.3404667

1. INTRODUCTION

Colorectal cancer is the second most encountered cancer type in women and the third most commonly occurring cancer type in men, with an average of 1.4 million new cases per year [1]. The estimated new colorectal cancer cases in 2019 are 145,600 and the estimated colorectal cancer-related deaths in 2018 were 51,020 [2]. Colorectal cancer will be a major health challenge in the U.S and the rest of the world if diagnosis and screening techniques are not improved. Timely diagnosis in colorectal allows the detection of malignant lesions that are yet to go beyond the submucosal layer [3].

NBI is a commonly used imaging technique in colonoscopy to diagnose cancerous legions [4]. NBI filters white light that is received by hemoglobin, a protein molecule in red blood cells that carries oxygen, so that it can be observed by a colonoscopy. While red light with a long wavelength is not absorbed, there are strong peaks at approximately 415 and 540 nm. To show details such as vessels, surface patterns, etc. in the mucosal layer, blue and green light should be used at 415 and 540 nm respectively.

NICE classification is based on three parameters; colors, vessels, and surface patterns[5]. According to NICE classification, there are 3 types: Type 1, Type 2 and Type 3. In Type 1 (Figure 1), the color of the lesion is the same or lighter than the background. In Type-1 there are not any vessels or if there are, they may be isolated lacy vessels. Also, the surface pattern has dark or white spots of uniform size, or there is a homogeneous absence of pattern. In Type 2 (Figure 2), the color is brown, relative to the background. Brown vessels are surrounding white structures, the surface pattern is oval, tubular or branched white structures surrounded by brown vessels. In Type 3, the color is brown to dark brown relative to the background, and sometimes there may

be patchy whiter areas. These kinds of lesions have disrupted or missing vessels. There is also an amorphous surface pattern. The long-term goal of this study is to design and develop a viable platform to automatically detect and classify polyps. Therefore, we performed a feasibility study to show the success rates of deep learning-based classification techniques to classify the colorectal tumor images according to Narrow-band imaging International Colorectal Endoscopic (NICE) classification [5] by using deep learning with convolutional neural networks (CNN) [6].

2. RELATED WORKS

Classifying a tumor is a crucial step of the colonoscopy. It helps to understand the characteristics of a tumor such as benign or malignant, widely spread or not, etc. For correct classification, different deep learning algorithms were developed in the previous years. Deep learning has become widespread in the last decade in the areas of image recognition and data science.

There are several methods developed for endoscopic image recognition such as Fisher-vector [7]–[9], Bag of Visual Words (BoVW) [10], Speeded Up Robust Features (SURF) [11], Scale-invariant feature transform (SIFT) [12] and Vector of Linearly Aggregated Descriptors (VLAD) [13], [14]. Nowadays, most of medical centers and hospitals are using the NBI systems in endoscopic examinations. However, it is difficult to classify tumors for medical students and novice surgeons in practice. Therefore, the development of computer-aided classification systems is increasing. Sonoyama et al. [15], tried to decrease computation cost without changing accuracy by using Fisher vector and VLAD. Fisher vector is a vectoral representation gathered by pooling image features with the use of Gaussian-Mixture-Model (GMM) [9], while VLAD uses k-mean to generate the features. Unlike VLAD, a Fisher vector stores feature-based second-order information which benefits classification performance [16]. On the other hand, Tamaki et al. [17], followed the BoVW approach with sampled SIFT features [18]and support vector machine (SVM) classifiers to classify 908 NBI images and had 94% to 96% accuracy. A staged filtering approach identifies scale-invariant features. The first stage identifies key locations in scale space by looking for locations of a difference-of-Gaussian function. A feature vector which explains the local image region sampled relative to its scale-space coordinate frame is generated by using each point. These vectors are called SIFT keys.

Deep learning is one of the neural network models which includes CNN models where layers are convolutions of their previous layers. The usage of CNN's is quickly increasing, and CNN's are replacing existing methods in the image recognition era due to its transcending performance. Fine-tuning from a pre-trained network is enough to get a high accuracy rate when there is a smaller dataset; however, to fully train a CNN, a large dataset is necessary [19]. Furthermore, features from CNN layers without any fine-tuning can be used for medical image recognition [20].

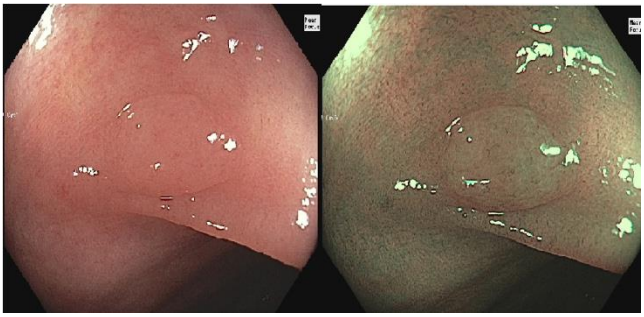

Figure 1. NICE 1 non-NBI and NBI tumor

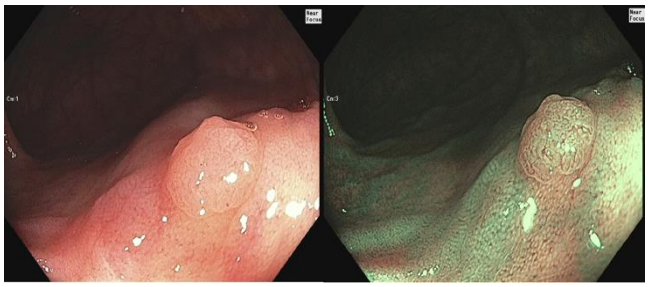

Figure 2. NICE 2 non-NBI and NBI tumor

3. METHODS

A CNN is a special class of neural networks that uses convolutional layers that convolve using the dot product of the pixels in an image. They take a weighted sum of a previously defined number of pixels called a filter in a certain region. In the initial layers of a CNN the network will learn based off parameters like edges, bright and dark spots etc. After multiple layers of filters, the CNN will begin recognizing full objects. In our case the CNN will begin to recognize and classify polyps.

CNN's are made up of convolutional layers which are layers that contain the previously mentioned filters. Pooling layers are used to concatenate and combine outputs from neurons which allows for larger portions of an image to be recognized. The last few layers in a CNN are called fully connected layers. These fully connected layers connect all the results in order to form a full classification of an image. These filters are further defined by activation functions. An activation function defines an output given an input or multiple input.

There is a great amount of image classification models that have millions of parameters such as GoogLeNet [21], AlexNet [22], InceptionV3 [23], etc. It will cost a lot of money and time to train them from scratch. Instead of doing this, we have used transfer learning which makes it easier by using a part of a model that has already been trained(pre-trained) on a new model. In this study, we used GoogLeNet, AlexNet, and InceptionV3, which are neural network architectures for image classification trained in the ImageNet Large-Scale Visual Recognition Challenge (ILSVRC) [22], [24]. The ImageNet dataset has 1000 different classes, 1,281,167 high-resolution images for training, 50,000 images for evaluation and 100,000 images for testing.

CNN architectures usually have a standard structure that consists of a convolutional layer followed by pooling and fully connected layers. AlexNet was firstly seen in the ImageNet LSVRC-2010. It was the first large scale CNN to perform well. The network has 60

million parameters and 650,000 neurons. The structure has eight layers which includes five convolutional layers, max-pooling layers, and three fully connected layers. In the ImageNet LSVRC-2012, they won the competition with 15.3% of the top-5 test error rate. On the other hand, GoogLeNet uses 12 times fewer parameters than the AlexNet, while it is significantly more accurate. It has 22 layers with parameters (27 layers if pooling layers are also counted). The total number of layers for the construction of the network is about 100. The exact number depends on the machine learning infrastructure which is used. The top-5 test error rate of GoogLeNet in ILSVRC 2014 was 7.89%. In the other model, InceptionV3 has 42 layers. Even when it is using less than 25 million parameters which is 5 times greater than GoogLeNet, the computation cost is only 2.5 higher than GoogleNet. The top-5 test error rate of InceptionV3 is 4.2% which is the lowest error rate regarding AlexNet and GoogLeNet. According to these results, we expect to get the most accurate results by using InceptionV3.

In this study, we used MATLAB 2018a, and one NVIDIA GTX960M graphic card v417.22. We have 2 classes: NICE 1 and NICE 2, and a total of 2604 images with a size of 64x64. We originally had 171 high-resolution images with a size of 1920x1080. To get more accurate results, we performed a preprocessing by removing reflections and cropping images to 64x64 samples manually (Figures 3 and 4). 90% of these images were used as the training set and 10% of them were used as the validation set.

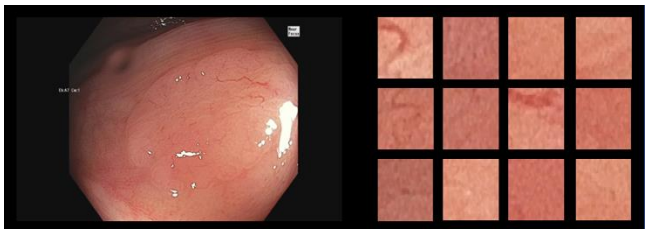

Figure 3. Cropping a NICE 1, 1920x1080 image to 64x64 samples.

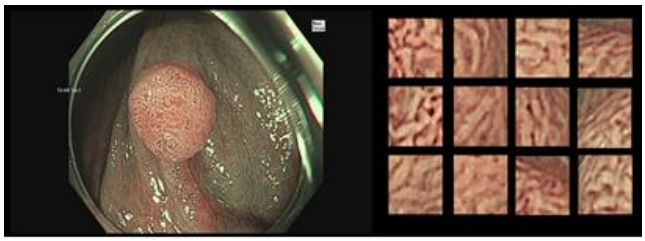

Figure 4. Cropping a NICE 2, 1920x1080 image to 64x64 samples.

4. RESULTS

We explain results with three different CNN models: AlexNet, GoogLeNet, and InceptionV3.

4.1 AlexNET

The first layer of the AlexNet, "data", requires 227x227x3 images. Therefore, we resized the 64x64 images to 227x227 (Figure 5). The next convolution layer applies 96 of 11x11x3 filter. After activation function and normalization layers, there is another layer which is a pooling layer that applies maximum pool by 3x3 filter. It repeats over and over, then reaches to the fully connected layers. Next two fully connected layers with 4096 nodes each. At the end, it has one fully connected layer with 1000 nodes, one SoftMax layer, and one output layer. The pretrained AlexNet model that we used has 25 layers. First, we extracted all layers, except the last three layers because the last three layers of the pre-trained network are configured for 1000 classes. However, we have only 2 classes for our comparison study, NICE 1 and NICE 2. Therefore, we kept the features from the early layers of the pre-trained network and retrained the last three layers. The average test result (Figure 6) of the 9 random executions is 88.19% (Min: 85.87%, Max: 91.82%).

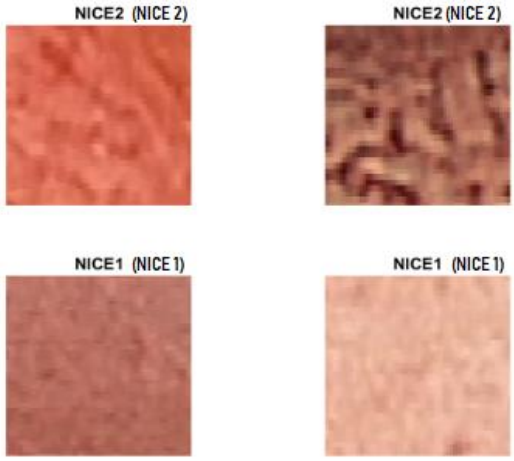

Figure 5. Sample validation images with their predicted labels and correct labels in parenthesis.

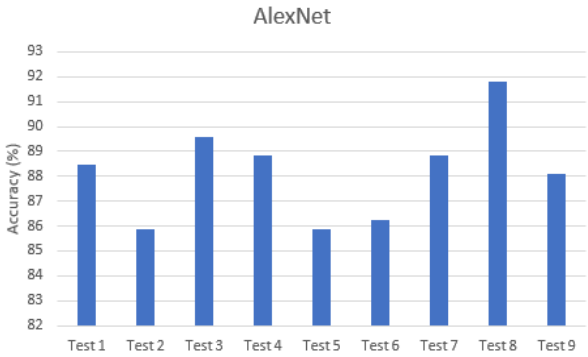

Figure 6. Test accuracies of AlexNet.

4.2 GoogLeNET

This architecture consists of 22 layers in deep. It reduces the number of parameters from 60 million (AlexNet) to 5 million. The pre-trained GoogLeNet model that we used has 144 layers. The first layer is the input layer which takes 224x224x3 images. Therefore, we resized the size of 64x64 images from the dataset to a size of 224x224. To retrain GoogLeNet, we replaced the last three layers which include information about the class labels and probabilities. After that, we extracted the connections and layers then set the learning rates to zero in the first 110 layers to prevent overfitting and speed up the network training. The average test result (Figure 7) of the 9 random executions is 85.73% (Min: 83.46%, Max: 88.46%).

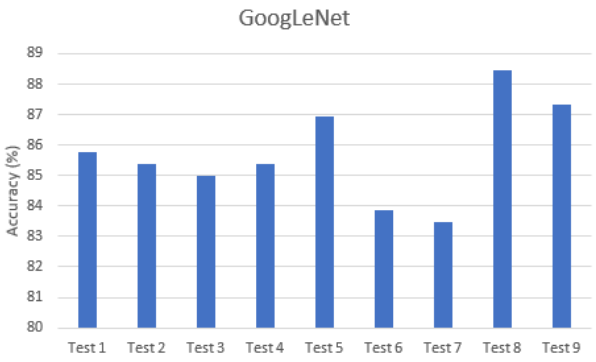

Figure 7. Test accuracies of GoogLeNet.

4.3 InceptionV3

The pre-trained InceptionV3 model that we used has 316 layers. Other than the other networks, the InceptionV3 model needs a size of 299x299 images. Therefore, we resized the size of 64x64 images to a size of 299x299 images. After that, the first convolution layer applies a 149x149x32 filter. Then, the normalization and activation layers are applied. In the first max-pooling layer the size of the filter is 73x73x64. To train the network for our classes, NICE 1 and NICE 2, we replaced the last three layers of the network. We also set the learning rates to zero in the first 110 layers the same as in the GoogLeNet. The average test result (Figure 8) of the 9 random executions is 92.39% (Min: 89.62%, Max: 95.77%).

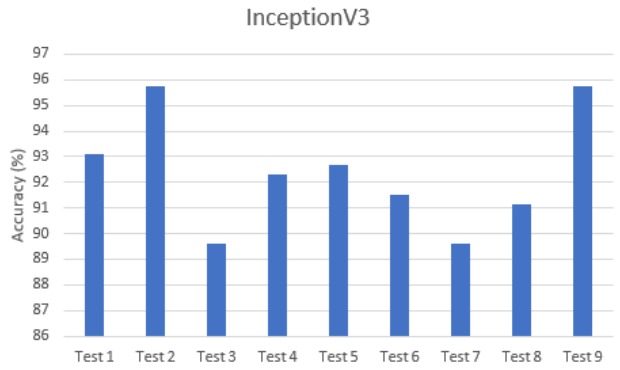

Figure 8. Test accuracies of InceptionV3.

5. CONCLUSIONS

We had 2 classes: NICE 1 and NICE 2, and a total of 171 high-quality colorectal tumor images. We performed a comparison study about the classification of the colorectal tumor images according to NICE classification by using deep learning models between GoogLeNet, AlexNet, and InceptionV3. Our results show that InceptionV3 has better accuracy with an average of 92.39%. We also achieved the most accurate result, 95.77%, with InceptionV3 model. We performed preprocessing by removing reflections and cropping images to 64x64x3 samples manually to improve accuracy. We were expecting that GoogLeNet was going the give better results than AlexNet according to their top-5 test error rates. However, in our transfer learning model infrastructure, AlexNet has an average 88.19% where GoogLeNet is 85.73%. This comparison study aims to show the accuracies of the deep learning models to classify the colorectal tumor images according to NICE classification. According to our study, these techniques will help surgeons to classify tumors and give promising results in the future. We would like to automate the preprocessing stage and improve this study to perform on real-time videos of the colonoscopy procedures for future work.

6. ACKNOWLEDGMENTS

This project was made possible by the Arkansas INBRE program, supported by a grant from the National Institute of General Medical Sciences, (NIGMS), P20 GM103429 from the National Institutes of Health (NIH). This project was also supported by NIH/NIAMS R44AR075481-01, NIH/NCI 5R01CA197491 and NIH/NHLBI NIH/NIBIB 1R01EB025241, R56EB026490.

7. REFERENCES

[1] "Colorectal cancer statistics," *World Cancer Research Fund*, Aug. 22, 2018. https://www.wcrf.org/dietandcancer/cancer-trends/colorectal-cancer-statistics (accessed Dec. 31, 2019).

[2] "Cancer of the Colon and Rectum - Cancer Stat Facts," *SEER*. https://seer.cancer.gov/statfacts/html/colorect.html (accessed Dec. 31, 2019).

[3] S. Kudo, H. Kashida, T. Nakajima, S. Tamura, and K. Nakajo, "Endoscopic diagnosis and treatment of early colorectal cancer," *World J. Surg.*, vol. 21, no. 7, pp. 694–701, 1997.

[4] K. Nonaka, M. Nishimura, and H. Kita, "Role of narrow band imaging in endoscopic submucosal dissection," *World J. Gastrointest. Endosc.*, vol. 4, no. 9, p. 387, 2012.

[5] N. Hayashi *et al.*, "Endoscopic prediction of deep submucosal invasive carcinoma: validation of the narrow-band imaging international colorectal endoscopic (NICE) classification," *Gastrointest. Endosc.*, vol. 78, no. 4, pp. 625–632, 2013.

[6] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[7] F. Perronnin and C. Dance, "Fisher kernels on visual vocabularies for image categorization," in *2007 IEEE conference on computer vision and pattern recognition*, 2007, pp. 1–8.

[8] F. Perronnin, J. Sánchez, and T. Mensink, "Improving the fisher kernel for large-scale image classification," in *European conference on computer vision*, 2010, pp. 143–156.

[9] J. Sánchez, F. Perronnin, T. Mensink, and J. Verbeek, "Image classification with the fisher vector: Theory and practice," *Int. J. Comput. Vis.*, vol. 105, no. 3, pp. 222–245, 2013.

[10] G. Csurka, C. Dance, L. Fan, J. Willamowski, and C. Bray, "Visual categorization with bags of keypoints," in *Workshop on statistical learning in computer vision, ECCV*, 2004, vol. 1, pp. 1–2.

[11] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool, "Speeded-up robust features (SURF)," *Comput. Vis. Image Underst.*, vol. 110, no. 3, pp. 346–359, 2008.

[12] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, 2004.

[13] H. Jégou, M. Douze, C. Schmid, and P. Pérez, "Aggregating local descriptors into a compact image representation," in *CVPR 2010-23rd IEEE Conference on Computer Vision & Pattern Recognition*, 2010, pp. 3304–3311.

[14] H. Jegou, F. Perronnin, M. Douze, J. Sánchez, P. Perez, and C. Schmid, "Aggregating local image descriptors into

compact codes," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 9, pp. 1704–1716, 2011.

[15] S. Sonoyama *et al.*, "Trade-off between speed and performance for colorectal endoscopic NBI image classification," in *Medical Imaging 2015: Image Processing*, 2015, vol. 9413, p. 94132D.

[16] M. Seeland, M. Rzanny, N. Alaqraa, J. Wäldchen, and P. Mäder, "Plant species classification using flower images—A comparative study of local feature representations," *PloS One*, vol. 12, no. 2, p. e0170629, 2017.

[17] T. Tamaki *et al.*, "Computer-aided colorectal tumor classification in NBI endoscopy using local features," *Med. Image Anal.*, vol. 17, no. 1, pp. 78–100, 2013.

[18] D. G. Lowe, "Object recognition from local scale-invariant features.," in *iccv*, 1999, vol. 99, pp. 1150–1157.

[19] T. Tamaki *et al.*, "Computer-aided colorectal tumor classification in NBI endoscopy using CNN features," *ArXiv Prepr. ArXiv160806709*, 2016.

[20] C.-K. Shie, C.-H. Chuang, C.-N. Chou, M.-H. Wu, and E. Y. Chang, "Transfer representation learning for medical image analysis," in *2015 37th annual international conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2015, pp. 711–714.

[21] C. Szegedy *et al.*, "Going deeper with convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1–9.

[22] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[23] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2818–2826.

[24] O. Russakovsky *et al.*, "Imagenet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, 2015.