2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)
Venice, Italy, April 8-11, 2019

HIGH ACCURACY PATCH-LEVEL CLASSIFICATION OF WIRELESS CAPSULE ENDOSCOPY IMAGES USING A CONVOLUTIONAL NEURAL NETWORK

Vinu Sankar Sadasivan[1] *and Chandra Sekhar Seelamantula*[2]*, Senior Member, IEEE*

[1]Department of Computer Science and Engineering, Indian Institute of Technology, Gandhinagar, India
[2]Department of Electrical Engineering, Indian Institute of Science, Bangalore, India
Email: vinu.sankar@iitgn.ac.in, chandra.sekhar@ieee.org

ABSTRACT

Wireless capsule endoscopy (WCE) is a technology used to record colored internal images of the gastrointestinal (GI) tract for the purpose of medical diagnosis. It transmits a large number of frames in a single examination cycle, which makes the process of analyzing and diagnosis of abnormalities extremely challenging and time-consuming. In this paper, we propose a technique to automate the abnormality detection in WCE images following a deep learning approach. The WCE images are split into patches and input to a convolutional neural network (CNN). A trained deep neural network is used to classify patches to be either *malign* or *benign*. The patches with abnormalities are marked on the WCE image output. We obtained an area under receiver-operating-characteristic curve (AUROC) value of about 98.65% on a publicly available test data containing nine abnormalities.

Index Terms— Gastrointestinal tract, classification, wireless capsule endoscopy, convolutional neural networks, deep learning

1. INTRODUCTION

Wireless capsule endoscopy (WCE), developed by Iddan et al. [1], is a painless way to diagnose the disease condition in the gastrointestinal (GI) tract. A capsule containing a miniature camera swallowed by the patient transmits color images of the GI tract through radio-telemetry at a high frame rate. A capsule endoscope (CE) gathers more data for diagnosing and it performs a simple scan that does not require a doctor to be present throughout the procedure. The number of images transmitted by a single CE is of the order of a few tens of thousands, which requires several hours of offline review by an expert to identify the abnormalities. Usually WCE image videos contain few frames with abnormalities and there are high chances for an expert to miss out on these frames.

Most of the computer-aided approaches for endoscopy diagnosis use either support vector machines or neural networks to perform classification on endoscopy data. Iakovidis et al. [2] considered color information and detected salient points in an endoscopy image using the SURF algorithm and constructed discriminative features capturing color information around the salient points. Convolutional Neural Networks (CNNs) have been found to be extremely helpful in extracting spatial features. Jia et al. [3] use a deep CNN to detect bleeding in GI tracts in WCE images. Sekuboyina et al. [4] developed a CNN architecture to extract features from WCE images and detect multiple abnormalities.

The present contribution is motivated by our previous research [4], wherein we used a CNN to classify eight types of abnormalities. In this work, we employ a CNN where the hyperparameters such as the learning rate, momentum, weight decay, batch size, dropout probability etc. have been tuned to improve the overall classification performance. We use the square of softmax cross-entropy with logits as the cost function and local response normalization as the normalizer for the input to each of the convolutional layer. We found that this improves convergence of the model and leads to superior performance. Our model achieves an overall maximum AUROC value of 98.65%, which is an increase of 15.5% over our previously reported results [4].

The WCE dataset we employ in our studies was originally recorded and shared by Iakovidis et al. [2]. The CE captures three images per second, each of dimension 320×320. These video frames were then classified by experts into broad categories such as vascular lesions, inflammatory lesions, lymphangiectasias, and polypoid lesions. The dataset consists of a total of 137 images. Of these, 77 images have abnormalities of 9 different types and 60 images correspond to no abnormality. Pixel-level annotations provided by experts are used as the ground truth for classification.

2. PROPOSED APPROACH

Texture and color are the two main characteristic features for abnormality detection in a WCE image. We draw our motivation for generating features based on texture and color information. In this work, we use a CNN to extract the color and texture information from the chromatic components taken

This work was sponsored by the Robert Bosch Centre for Cyberphysical Systems.

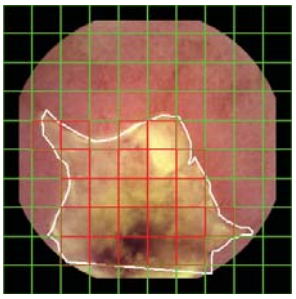


Fig. 1: The patches labeled *malign* are marked red and those labelled *benign* are marked green. The white contour shows the malign region, as given in the annotated ground truth.

from the CIE-*Lab* space of the images. Our algorithm aims at classifying a patch in the frame belonging to an abnormality (*malign* patch) or not (*benign* patch).

2.1. Preprocessing

All the images in the dataset are converted into the CIE-*Lab* space. This decision is based on our previous work [4], which showed that it aids in extracting more general features such as color and texture. The images are further divided into patches of size 36×36 (cf. Fig. 1), resulting in 100 patches per image; thereby, giving us an adequate number of data points to train the CNN model. There are 137 images in total, which gives us a total of 13700 patches. However, we note that, the dataset is skewed, in the sense that the ratio of *malign* to *benign* patches in the dataset is very low. Hence, training the model with the given dataset will bias it towards predicting a patch to be *benign*. To avoid this, the dataset is expanded using data augmentation techniques, such as rotation and reflection – maintaining the ratio between the *malign* and *benign* examples. These data augmentation steps would also result in invariances to rotation and reflection as the capsule imaging is indeed prone to such effects. A patch is labeled to be *benign* if less than 50% of the pixels on it are benign compared to its annotation. The new expanded dataset is randomly divided into training and testing subsets, maintaining the distribution of data in each class. The training dataset that we chose is about 0.90 times the overall size of the augmented dataset.

2.2. Training and CNN Architecture

The input to the CNN model is the $36 \times 36 \times 3$ (height $\times$ width $\times$ #channels) normalized patch. The 320×320 WCE images are zero-padded to make them 360×360 images. The patches are extracted from these padded-images as it helps the model to learn the background as well (cf. Fig. 1). Normalization bring all pixels (across patches) to the same scale, and furthermore, improves convergence. An overview of the proposed CNN architecture is shown in Fig. 2. We use three kernels with a planar spread of 5×5 in all convolutional layers with a stride of 1 and max-pooling layers of kernel size 2×2 with a stride of 2. A local response normalization (LRN) [5] is performed over the tensor, which is the input to the max-pooling layer. Normalization speeds up the convergence of the model and dampens the responses that are uniformly large in any given local neighborhood. We employ a rectified linear unit (ReLU) activation in both convolutional and fully-connected layers except the final one (which converts the scores into probabilities for each class). The multi-layer perceptron network of structure 75-10-2 is used for the classification task. A dropout keep probability of 0.90 to maintain each element is applied to the last fully connected hidden layer to regularize the network. Due to inherent resistance to overfitting by virtue of shared weights across the image, regularization is not employed in convolutional layers. One-hot encoding is used to label the output classes. All the weights and biases are initially set with a value of 0.30, which showed an improvement in convergence of the model. The loss function L employed is the square of softmax cross-entropy with logits. Logits are preferred in categorization tasks [6, 7]. The learning rate is set to be 0.001 with β_1 value set to be 0.9 and β_2 value set to be 0.999 for the optimizer, where β_1 and β_2 are the exponential decay rates for the first and second moment estimates, respectively. The batch size is set to 64. The maximum number of epochs used for training the model is 50. Early stopping is employed to avoid overfitting. The i^{th} convolutional layer, and the softmax classsifier of the architecture can be mathematically modelled as follows,

$$y_i = f(\tilde{X}_i; \theta_i) \tag{1}$$

$$a_k = \frac{e^{z_k}}{\sum_{\forall j} e^{z_j}} \tag{2}$$

where y_i, and f are the output, and the operations, respectively, over LRN of input X_i to the network, notated as $\tilde{X}_i$, with parameters θ_i. Activation and output for logit k [6] from the network fed to the classifier is notated as a_k and z_k, respectively. The abnormal patches are highlighted as shown in Fig. 1. The following metrics are used for performance assessment: sensitivity (SN), specificity (SP) [8], and area under receiver-operating-characteristic curve (AUROC) [9]. The model's performance was measured by evaluating 10 times on randomly sampled training and test sets. The mean values of the metrics are displayed in Table 2.

3. RESULTS

The metric values are calculated for classification of each disease individually and for the whole dataset. A 10-fold Monte Carlo cross-validation is used to compare the proposed model with the state-of-the-art methods (as reported in [2, 4]) as highlighted in the Table 1. The overall maximum AUROC, SN, and SP values on the test data achieved

Table 1: A comparison of the AUROC (area under receiver-operating-characteristic curve) values calculated for the classification of each disease individually with the state-of-the-art approaches [2, 4].

Disease	**Proposed** (in %)	**Sekuboyina et al.** [4] (in %)	**Iakovidis et al.** [2] (in %)
Angiectasia	**94.42 ± 1.54**	NA	NA
Apthae	**94.32 ± 2.28**	78.81 ± 10.14	79.1 ± 13.1
Bleeding	**94.45 ± 0.47**	64.08 ± 5.39	83.5 ± 10.1
Chylous cysts	**95.68 ± 0.45**	87.85 ± 6.8	87.6 ± 4.3
Lymphangiectasias	**96.90 ± 0.50**	95.95 ± 2.28	96.3 ± 3.6
Polypoids	**92.36 ± 2.48**	73.86 ± 7.11	85.9 ± 6
Stenoses	**96.35 ± 0.54**	76.73 ± 3.65	80.2 ± 13.4
Ulcers	**94.35 ± 2.49**	89.4 ± 2.26	76.2 ± 10
Villous Oedema	**95.32 ± 1.58**	78.38 ± 7.38	92.3 ± 7.6

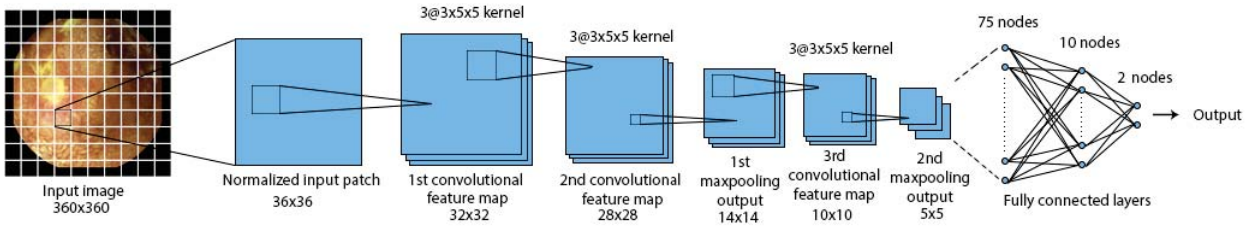
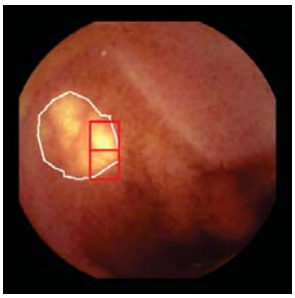


Fig. 2: The CNN architecture used to train and test the model (after [4]). A detailed description of the network is given in Section 2.

(a) Abnormality detected in Angiectasia image



(b) Abnormality detected in Bleeding image

Fig. 3: The figure shows a successful (left), and an unsuccessful (right) classification of the proposed method. Here, the red patches are classified as malignant by the model.

are 98.65%, 96.64%, and 98.77%, respectively. The mean values of measured metrics after cross-validation are in Table 2. The AUROC of the proposed CNN model outperforms the state-of-the-art methods [2, 4] in detecting all the 9 abnormalities. Fig. 3(a) shows the patches corresponding to a chylous cyst being detected successfully by the proposed technique. Fig. 3(b) shows a case where the model fails to detect a bleeding image correctly. A plausible reason for the failure in bleeding detection, as in Fig. 3(b), might be due to the lack of prominent texture patterns, from which the CNN could benefit.

Table 2: Mean values of performance measure of the model.

Disease	**SN** (in %)	**SP** (in %)	**AUROC** (in %)
Angiectasia	95.24	96.67	94.42
Apthae	93.64	99.56	94.32
Bleeding	95.23	94.56	94.45
Chylous Cysts	93.75	98.51	95.68
Lymphangiectasias	97.39	98.51	96.90
Polypoids	95.67	94.01	92.36
Stenoses	94.70	99.09	96.35
Ulcers	94.75	98.93	94.35
Villous Oedema	94.05	99.75	95.32
Overall	**94.95**	**97.72**	**95.36**

4. CONCLUSIONS

We addressed the problem of automatic abnormality detection in WCE image videos by taking our previously proposed CNN architecture as a starting point. We showed that fine-tuning the hyperparameters boosted the AUROC by 15.5%. The validations were carried out on a publicly available WCE image dataset that contains images of 9 abnormalities apart from normal images. The proposed model is good at classifying abnormal patches from normal patches. The work could be extended by carrying out the validations on a much larger dataset and fine-tuning the network architecture and parameters.

5. REFERENCES

[1] G. J. Iddan, G. Meron, A. Glukhovsky, and P. Swain, "Wireless capsule endoscopy," *Nature*, vol. 405, pp. 417–417, 2000.

[2] D. K. Iakovidis and A. Koulaouzidis, "Automatic lesion detection in wireless capsule endoscopy – a simple solution for a complex problem," in *Proceedings IEEE International Conference on Image Processing*, Oct 2014, pp. 2236–2240.

[3] X. Jia and M. Q. H. Meng, "A deep convolutional neural network for bleeding detection in wireless capsule endoscopy images," in *Proceedings International Conference of the IEEE Engineering in Medicine and Biology Society*, Aug 2016, pp. 639–642.

[4] A. K. Sekuboyina, S. T. Devarakonda, and C. S. Seelamantula, "A convolutional neural network approach for abnormality detection in wireless capsule endoscopy," in *Proceedings IEEE International Symposium on Biomedical Imaging*, Apr. 2017, pp. 1057–1060.

[5] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proceedings International Conference on Neural Information Processing Systems - Volume 1*, USA, 2012, NIPS'12, pp. 1097–1105, Curran Associates Inc.

[6] S. Hasegawa and T. I. Kurita, "Face and non-face classification by multinomial logit model and kernel feature compound vectors," in *Proceedings Neural Information Processing Systems*, Nov 2002, vol. 2, pp. 996–1000 vol. 2.

[7] Y. Hai, K.-L. Tsui, and M. J. Zuo, "Gear crack level classification based on multinomial logit model and cumulative link model," in *Proceedings IEEE Prognostics and System Health Management Conference*, May 2012, pp. 1–6.

[8] A. G. Lalkhen and A. McCluskey, "Clinical tests: sensitivity and specificity," *Continuing Education in Anaesthesia Critical Care & Pain*, vol. 8, no. 6, pp. 221–223, 2008.

[9] Andrew P. Bradley, "The use of the area under the ROC curve in the evaluation of machine learning algorithms," *Pattern Recognition*, vol. 30, no. 7, pp. 1145 – 1159, 1997.