The Medico-Task 2018: Disease Detection in the Gastrointestinal Tract using Global Features and Deep Learning

Vajira Thambawita[1,3], Debesh Jha[1,4], Michael Riegler[1,3,5], Pål Halvorsen[1,3,5],
Hugo Lewi Hammer[2], Håvard D. Johansen[4], and Dag Johansen[4]
[1]Simula Research Laboratory, Norway [2]Oslo Metropolitan University, Norway [3]Simula Metropolitan, Norway
[4]University of Tromsø, Norway [5]University of Oslo, Norway
Contact:vajira@simula.no,debesh@simula.no

ABSTRACT

In this paper, we present our approach for the 2018 Medico Task classifying diseases in the gastrointestinal tract. We have proposed a system based on global features and deep neural networks. The best approach combines two neural networks, and the reproducible experimental results signify the efficiency of the proposed model with an accuracy rate of 95.80%, a precision of 95.87%, and an F1-score of 95.80%.

1 INTRODUCTION

Our main goal for the Medico Task [15] is to classify findings in images from the Gastrointestinal (GI) tract. This task provides two types of input data: Global Features (GFs) and original images. The 2017 Medico Task consisted of a balanced dataset with only 8 classes [12] whereas the current task consists of a highly imbalanced dataset with 16 classes [11, 12], i.e., making this years task more complicated. Different approaches have been used in the last year medico task [5, 7, 9, 10, 14, 17] based on GFs extractions and Convolutional Neural Networks (CNN) methods. We extend upon these solutions and present our solutions based on both GFs and transfer learning mechanisms using CNN. We achieve best results combining two CNNs and using an extra multilayer perceptron to combine the outputs of the two networks.

2 APPROACHES

We approach the problem of GI tract disease detection with small training datasets using five different methods: two based on GF extractions, and three based on CNN with transfer learning described below.

2.1 Global-feature-based approaches

Method 1 and **Method 2** use the concept of GFs. For the extraction of GFs, we use Lucence Image Retrieveal (LIRE) [6]. GFs are easy and fast to calculate, and can also be used for image comparison, image collection search and distance computing [14]. Based on [13, 16], we use Joint Composite feature (JCD), Tamura, Color layout, Edge Histogram, Auto Color Correlogram and Pyramid Histogram of Oriented Gradients (PHOG). These features represent the overall properties of the images. Adding more GFs is possible, but it may increase the redundant information which can reduce the overall classification performance.

The extracted features are sent to the different machine learning classifier for the multi-class classification. **Method 1** makes the use of extracted GFs that are sent to SimpleLogistic (SL) classifier. We input the same selected set of features to the logistic model tree (LMT) classifier in **Method 2**.

2.2 Transfer learning based approaches

Our CNN approaches use transfer learning mechanism with pre-trained models using the ImageNet dataset [18]. Resnet-152 [3] and Densenet-161 [4] have been selected, and this selection is based on top 1-error and top-5-errors rate of pre-trained networks in the Pytorch [8] deep learning framework.

One of the main problems of the given dataset is the "out of patient"-category which has only four images while other classes have a considerable number. The colour distribution of this class shows a completely different colour domain compared to the other categories. We identified this difference via manual investigations of the dataset and moved all four images of this category into the corresponding validation set folder. Then, the training set folder is filled with random Google images which are not related to the GI tract. To overcome the problems of stopping training in a local minima, we use the stochastic gradient descent [1] method with dynamic learning rate scheduling. The losses (loss 1 and loss 2 in Figure 1) of CNN methods were calculated for each network separately. Additionally, horizontal flips, vertical flips, rotations and re-sizing data augmentations have been applied to overcome the problem of over-fitting.

Method 3 uses transfer learning with Resnet-152 which has the top-1-error and top-5-error rates. The last fully connected layer of Resnet-152, which is originally designed to classify 1000 classes of the ImageNet dataset, has been changed to classify the 16 classes in the MEdico task. Usually, the transfer learning freezes pre-trained layers to avoid back propagation of large errors. This is because of newly added layers with random weights. However, we did not freeze the pre-trained layers, because modifying only the last layer cannot propagate huge errors backwards in transfer learning. The network was trained until it reached to the maximum validation accuracy of the validation dataset.

Method 4 extends Method 3 by using two parallel pre-trained models, Resnet-152 and Densenet-161, to get a cumulative decision at the end as depicted in Figure 1. The classification is based on an average of the two output probability vectors. Finally, one loss value was calculated and propagated for updating weights. However, this yields a restriction of updating weights of networks Resnet-152 and Densenet-161 separately as they required. Therefore, we calculated two different loss values (loss 1 and loss 2 in Figure 1) from each network to update their weights separately. Both

MediaEval'18, 29-31 October 2018, Sophia Antipolis, France

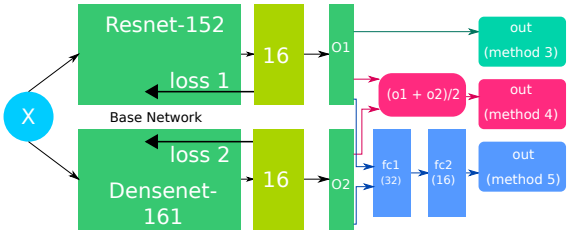

Figure 1: Block diagram of the CNN methods

networks were trained simultaneously until it reached to the best validation accuracy by changing hyper-parameters manually.

Method 5 was constructed to overcome the limitation of calculating the average of the probabilistic output of the two networks used in Method 4. Instead of calculating the average using the simple mathematical formula, another multilayer perceptron (MLP) has been merged with the above network to identify complex mathematical formula to get the cumulative decision as illustrated in Figure 1. Therefore, we passed the probability output of two networks (16 probabilities from each network) to a new MLP with 32 inputs, 16 outputs (via sigmoid layer) and one hidden layer with 32 units. In this, we used pre-trained Resnet-152 and Densenet-161 using the dataset and froze them before training the MLP. Then, we trained only the MLP to identify the best mathematical formula to get the cumulative decision.

3 RESULTS AND ANALYSIS

We have divided the development dataset into a training set (70%) and a validation set (30%). For the GFs based approach, ensembles of six extracted GFs were fetched to all the available machine learning classifiers (with different parameters) using WEKA[2] library. The SL and LMT classifiers outperform all other available classifiers for the dataset. The other promising classifier were Sequential minimal optimization (RBF kernel), and a combination of PCA with LibSVM (RBF) classifier.

On validation set, all the CNN methods (3-5) show accuracies of around 95% and specificities of around 99%. These are always better than the GFs based extraction methods (1,2) which have accuracies of around 82% and specificities of around 98%. According to the task organizers' evaluation results of the test dataset, Methods 3 to 5 show accuracies and specificities of around 99% again,which demonstrates our CNN methods are not overfitted with validation dataset.

Method 5 and 4 with Resnet-152 and Densenet-161 performs better compared to the Method 3 which has only Resnet-152 because of the capability of deciding the final answer based on two answers generated from two deep learning networks. However, getting a cumulative decision based on simple averaging function (Method 4) shows poor performance than the decision taken from a MLP (Method 5). As a result, Method 5 shows better results than method 4 by increasing the accuracy from 0.955 to 0.958. Therefore, Method 5 has been selected as our best method and confusion matrix represented in Table 1 was generated. An overview of the individual results obtained from five different experiments along with their performance metrics is presented in Table 2. Results obtained from the organizers for the test dataset is presented in the Table 3.

Table 1: The Confusion Matrix of Method 5 in our study

A:blurry-nothing, **B**:colon-clear, **C**:dyed-lifted-polyps, **D**:dyed-resection-margins, **E**:esophagitis,**F**:instruments, **G**:normal-cecum, **H**:normal-pylorus, **I**:normal-z-line, **J**:out-of-patient, **K**:polyps, **L**:retroflex-rectum, **M**:retroflex-stomach, **N**:stool-inclusions, **O**:stool-plenty, **P**:ulcerative-colitis

Actual class \ Predicted class	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
A	53															
B		81														
C			130	7												1
D			3	122												
E					115				19							
F						10					1					
G							125									
H								132								
I					11				121							
J						1				3						
K		1					6	2			172					
L							1					71				
M												2	118			
N														39		
O															110	
P					1	1	2				4	1				129

Table 2: Validation results

Method	REC	PREC	SPEC	ACC	MCC	F1	FPS
1	0.855	0.793	0.989	0.816	0.814	0.823	79
2	0.816	0.817	0.984	0.816	0.800	0.815	12
3	0.9536	0.9543	0.9968	0.9536	0.9498	0.9535	64
4	0.9555	0.9563	0.9969	0.9555	0.9519	0.9554	29
5	0.9580	0.9587	0.9971	0.9580	0.9546	0.9580	29

Table 3: Official results

Method	REC	PREC	SPEC	ACC	MCC	F1
1	0.8457	0.8457	0.9897	0.9807	0.8353	0.8456
2	0.8457	0.8457	0.9897	0.9807	0.8350	0.8457
3	0.9376	0.9376	0.9958	0.9922	0.9335	0.9376
4	0.9400	0.9400	0.9960	0.9925	0.9360	0.9400
5	0.9458	0.9458	0.9964	0.9932	0.9421	0.9458

The main considerable point in the confusion matrix in Table 1 is misclassification between categories E: esophagitis and I: normal-z-line. A large number of misclassifications like 30 images from the validation set occurred and a manual investigation was done to identify the reason. We notice that the images of these two categories were very similar to each other because of the close location in the GI tract, and identifying these is also a challeng for physicians.

4 CONCLUSION

In this paper, we presented five different methods for the multi-class classification of GI tract diseases. The proposed approach are based on the GFs, and pre-trained CNN with transfer learning mechanism. The combination of Resnet-152 and Densenet-161 with an additional MLP achieved the highest performance with both the validation dataset and the test dataset provided by the task organizers. We show that a combination of pre-trained deep neural models on ImageNet has better capabilities to classify images into the correct classes because of cumulative decision-making capabilities. For future work, we will combine deeper CNNs parallelly to add more cumulative decision taking capabilities for classifying multi-class objects. In addition to that, Generative Adversarial Network (GAN) methods can be utilized to handle imbalance dataset by generating more data to train deep neural networks.

REFERENCES

[1] Ian Goodfellow, Yoshua Bengio, Aaron Courville, and Yoshua Bengio. 2016. *Deep learning*. Vol. 1. MIT press Cambridge.

[2] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H Witten. 2009. The WEKA data mining software: an update. *ACM SIGKDD explorations newsletter (SIGKDD Explor. Newsl.)* 11, 1 (2009), 10–18.

[3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*. 770–778.

[4] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. 2017. Densely Connected Convolutional Networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2261–2269.

[5] Yang Liu, Zhonglei Gu, and William K Cheung. 2017. HKBU at MediaEval 2017 Medico: Medical multimedia task. In *Working Notes Proceedings of the MediaEval 2017 Workshop (MediaEval 2017)*.

[6] Mathias Lux, Michael Riegler, Pål Halvorsen, Konstantin Pogorelov, and Nektarios Anagnostopoulos. 2016. LIRE: open source visual information retrieval. In *Proceedings of the 7th International Conference on Multimedia Systems (MMSys)*. ACM, 30.

[7] Syed Sadiq Ali Naqvi, Shees Nadeem, Muhammad Zaid, and Muhammad Atif Tahir. 2017. Ensemble of Texture Features for Finding Abnormalities in the Gastro-Intestinal Tract. *Working Notes Proceedings of the MediaEval 2017 Workshop (MediaEval 2017)*.

[8] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. 2017. Automatic differentiation in PyTorch. In *Proceedings of 31st Conference on Neural Information Processing Systems (NIPS)*.

[9] Stefan Petscharnig and Klaus Schöffmann. 2018. Learning laparoscopic video shot classification for gynecological surgery. *An International Journal of Multimedia Tools and Applications* 77, 7 (2018), 8061–8079.

[10] Stefan Petscharnig, Klaus Schöffmann, and Mathias Lux. 2017. An Inception-like CNN Architecture for GI Disease and Anatomical Landmark Classification. In *Working Notes Proceedings of the MediaEval 2017 Workshop (MediaEval 2017)*.

[11] Konstantin Pogorelov, Kristin Ranheim Randel, Thomas de Lange, Sigrun Losada Eskeland, Carsten Griwodz, Dag Johansen, Concetto Spampinato, Mario Taschwer, Mathias Lux, Peter Thelin Schmidt, Michael Riegler, and Pål Halvorsen. 2017. Nerthus: A Bowel Preparation Quality Video Dataset. In *Proceedings of the 8th ACM on Multimedia Systems Conference (MMSYS)*. ACM, 170–174.

[12] Konstantin Pogorelov, Kristin Ranheim Randel, Carsten Griwodz, Sigrun Losada Eskeland, Thomas de Lange, Dag Johansen, Concetto Spampinato, Duc-Tien Dang-Nguyen, Mathias Lux, Peter Thelin Schmidt, and others. 2017. Kvasir: A multi-class image dataset for computer aided gastrointestinal disease detection. In *Proceedings of the 8th ACM on Multimedia Systems Conference (MMSYS)*. ACM, 164–169.

[13] Konstantin Pogorelov, Michael Riegler, Sigrun Losada Eskeland, Thomas de Lange, Dag Johansen, Carsten Griwodz, Peter Thelin Schmidt, and Pål Halvorsen. 2017. Efficient disease detection in gastrointestinal videos–global features versus neural networks. *An International Journal Multimedia Tools and Applications* 76, 21 (2017), 22493–22525.

[14] Konstantin Pogorelov, Michael Riegler, Pål Halvorsen, Carsten Griwodz, Thomas de Lange, Kristin Ranheim Randel, Sigrun Eskeland, Dang Nguyen, Duc Tien, Olga Ostroukhova, and others. 2017. A comparison of deep learning with global features for gastrointestinal disease detection. In *Working Notes Proceedings of the MediaEval 2017 Workshop (MediaEval 2017)*.

[15] Konstantin Pogorelov, Michael Riegler, Pål Halvorsen, Thomas De Lange, Kristin Ranheim Randel, Duc-Tien Dang-Nguyen, Mathias Lux, and Olga Ostroukhova. 2018. Medico Multimedia Task at MediaEval 2018. In *Working Notes Proceedings of the MediaEval 2018 Workshop*.

[16] Michael Riegler, Konstantin Pogorelov, Sigrun Losada Eskeland, Peter Thelin Schmidt, Zeno Albisser, Dag Johansen, Carsten Griwodz, Pål Halvorsen, and Thomas De Lange. 2017. From annotation to computer-aided diagnosis: Detailed evaluation of a medical multimedia system. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* 13, 3 (2017), 26.

[17] Michael Riegler, Konstantin Pogorelov, Pål Halvorsen, Carsten Griwodz, Thomas Lange, Kristin Ranheim Randel, Sigrun Eskeland, Dang Nguyen, Duc Tien, Mathias Lux, and others. 2017. Multimedia for medicine: the medico Task at mediaEval 2017. In *Working Notes Proceedings of the MediaEval 2017 Workshop (MediaEval 2017)*.

[18] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, and others. 2015. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)* (2015).