Image Quality Assessment of Cytology Images using Deep Learning

Tomé Albuquerque[1,2]
tome.m.albuquerque@inesctec.pt

Maria João M. Vasconcelos[3]
maria.vasconcelos@fraunhofer.pt

Jaime S. Cardoso[1,2]
jaime.cardoso@inesctec.pt

[1] Faculty of Engineering of the University of Porto
Porto, Portugal

[2] INESC TEC
Porto, Portugal

[3] Fraunhofer Portugal AICOS
Porto, Portugal

Abstract

The massive growth of digital color image contents in medical field, due to the spread of advanced multimedia devices capable of acquisition, transmission, and storage of digital data, demands improvements in image quality assessment (IQA) methods. Cervical cancer ranks as the fourth most common cancer among females worldwide with roughly 528,000 new cases yearly. Significant progress in the realm of artificial intelligence, particularly in neural networks and deep learning, helps physicians to classify cervical cancer more accurately using cytology and colposcopy images. However, it is necessary to ensure good image quality for decent performance of classifying methods. In this paper, we address a binary IQA problem (bad versus good quality) of cytology images with three different widely used architectures: VGG16, MobileNet, and ResNet50. The experimental results show the good performance of deep learning algorithms for IQA.

1 Introduction

Medical image quality assessment plays an important role not only in the design and manufacturing processes of image acquisition but also in the optimization of decision systems. Cervical cancer remains the fourth most common cause of cancer death in women worldwide. Despite the outburst of recent scientific advances to find an effective treatment, there is no effective method, especially when diagnosed in an advanced stage. However, screening tests such as cytology or colposcopy, have been responsible for a strong decrease in cervical cancer deaths.

Cytology microscope images need high-level microscopic magnification for a consistent characterization, but it is necessary to preserve an appropriate image quality [1]. In most of the cases, the cells in test slides are frequently spread in a multi-layer way which raises a challenge for a good focusing. Thus, it is necessary to use different focus levels for correct digital representation with good image quality. Powerful auto-focusing techniques using IQA methods are used in automated microscopy to prevent the loss of image quality [2]. The adequacy assessment of the cytology image has been studied by several researchers which propose different approaches to guarantee the The Bethesda System (TBS) minimum criteria (cellularity, obscuring factors, and the evidence of transformation zone) [3]. Most of the works on literature discard the importance of IQA and are focused on classification problems in cytology. Therefore, due to the lack of scientific work regarding the IQA on cervical cancer screening method, it is necessary to fulfil that gap with more research on this field to improve the actual state-of-the-art.

1.1 Brief summary

In this work, some approaches for non-reference image quality assessment (NR-IQA) are presented using feature extraction and learning. Thus, different convolutional neural network (CNN) based models, pre-trained on ImageNet dataset, were used and fine-tuned to predict the quality score value of several images. The first models uses three different architectures VGG16, MobileNet, and ResNet50) to predict the image quality of a blood cells microscopy dataset, labelled in a multiclass problem (4 classes).

Due to the lack of annotated cytology dataset regarding image quality, it was used as reference a microscopic blood cell sample dataset. After selecting the best model, the weights of the best model were used to initiate the train of a new model to classify IQA on a new IQA dataset created in this work with cytology microscope images. This new dataset contains reference images and distorted images obtained from the reference images. The classification of IQA on the cytology dataset is done on a binary problem (bad quality vs good quality), this classifier intends to learn low-notion quality features by distinguish between original/ reference images and distorted images.

2 Methods

2.1 Dataset

In this work, two different datasets were used in the train of IQA models. The microscope slide preparations of the first dataset, from blood samples, were obtained in Centro Hospitalar de São João (Porto), and digitalized by Fraunhofer AICOS (FhP-AICOS) within the scope of MpDS project [1] using μsmartscope. The blood cell data was annotated by doctors and contain a total of 1854 images divided into 4 different classes according to quality score of the image, which can be bad, fair, good, or excellent quality. The second dataset was collected in Hospital Professor Doutor Fernando Fonseca (Lisbon) within the scope of CLARE project using an updated version of μsmartscope, adapted to acquire samples with 400 times magnification the images were also acquired as the first dataset by FhP-AICOS using μsmartscope. This second dataset consists of 4088 images of microscope pap smear slide preparations (liquid-based citology samples), in which 817 are reference images and 3271 are images generated from the reference images with four different types of distortions (Gaussian noise, blur, salt and pepper noise and speckle noise). Thus, for every image of reference four new images were created with the distortions mentioned before as showed in Figure 1. This new cytology dataset is divided in 2 different classes, distorted images (bad quality images), and reference images (good quality images).

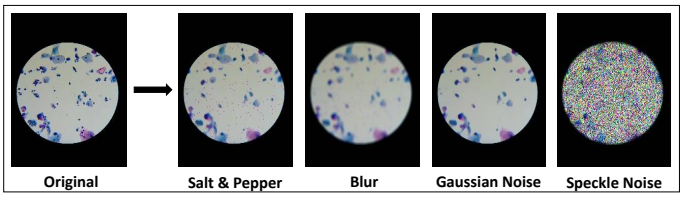

Figure 1: Example of the four different types distortions applied on the original images of cytology dataset.

2.2 Data Pre-processing

First, is worth mentioning that both datasets were divided into train, validation, and test subsets (60/20/20 %) in 5 different k-folds for cross-validation, maintaining the ratio among the different classes. The acquisition of the images in both datasets was done by different experts. In all the images the ROI is a circular region but, in some cases, that region is not in the centre of the image. To guarantee that the input image for the CNN models is a perfect square of the ROI it was necessary do an automatic crop at the limits of the circle creating a square shape image. The images were also resized to 224 x 224 and normalized from the pixel values range of 0-255 to the range 0-1 to speed up the train. Both datasets used in this work had a small number of labelled images (1854/4088), to overcome this obstacle it was used data augmentation. Therefore, during the training of the models a series of random transformations have been done in each training epoch for every single image. The transformations

[1] *https://www.aicos.fraunhofer.pt/en/our_work/portfolio/micron.html*

applied included image rotation with a range of 90°, width and height shift, horizontal flip, zoom (in or out).

2.3 Convolutional Neural Networks

Convolutional neural network (CNN) is a class of deep learning algorithms that are organized in several connected successive stages, specifically convolutional (conv), pooling (pool) layers, and activation functions. CNN has learnable parameters named weights that can be updated using several matrix multiplications and its goal is to reduce the images into a form that is easier to process and classify. After the convolutional block of the neural network, it follows a Fully Connected (FC) network that has as input the flatten feature map with origin in the convolutional output. The last layer of the FC computes the classification probability for each class using SoftMax regression for multiclass classification or sigmoid for binary classification.

3 Experimental Details

Two different models were trained and tested in this work. The first model 1 was tested with three different CNN architectures. All the architectures were directly fed with the blood cells and cytology images. The CNN architecture that achieves the best results in the IQA of blood cells dataset was used on model 2 for cytology IQA. The tested CNN architectures are the following: VGG16, MobileNet and ResNet50. For multiclass IQA of the blood cells, it was implemented the model 1 with the 3 different convolutional architectures mentioned above. Model 2 follows the same pattern of model 1, however, it is a CNN model for binary classification. This model classifies the image quality of cytology cells in two different classes. The pre-trained weights of model 1 are used to initiate the train of model 2 through transfer learning to increase the performance and diminish training time. The hyperparameters were adjusted just for model 1 with the grid search where the hyperparameters tuned were the learning rate (LR $\in$ [0.001; 0.0001]), the batch size (BS $\in$ [16; 32]), and the dropout rate (Dt $\in$ [0.2; 0.5]). During 500 epochs (using ModelCheckPoint and EarlyStopping TensorFlow callbacks) the model 1 was tested with all these hyperparameters. The losses used in this work were categorical and binary cross-entropy for multiclass and binary classification respectively.

4 Results and Discussion

The best combination of the hyperparameters found for model 1 after a fine-tuning was LR of 0.0001, BS of 32, and Dt of 0.2. These hyperparameters were chosen concerning the validation subset. The best results after the grid search for the first model (model 1 - multiclass) are represented in Table 1. For the multiclass problem, VGG16 achieved the best results with higher values in all the presented metrics when compared with MobileNet and ResNet50.

Table 1: Results of model 1 for multiclass classification task of IQA in blood cells test subset.

	Accuracy (%)	**Precision (%)**	**recall (%)**	**AUC (%)**
MobileNet	76.05± 1.74	76.34 ± 2.75	76.05 ± 1.74	86.85 ± 1.39
VGG16	**78.91 ± 1.97**	**79.16 ± 2.17**	**78.91 ± 1.97**	**87.64 ± 1.82**
ResNet50	76.97 ± 2.39	77.29 ± 2.67	76.97 ± 2.39	83.06 ± 3.38

To confirm if the resize of the images to 224 x 224 did not contribute to loss of information an additional model trained with resized images but with bigger dimensions (512 x 512) was created and tested. Due to the imbalance in the number of images per class in blood cells dataset classes a new model was done by oversampling the minority classes. The results for these different approaches using different shows no upgrades in the metrics. The model 1 trained with VGG16 as the convolutional block and using as input 224 x 244 resized images achieved the best performance. Thus, the pre-trained weights of this model 1 will be used by model 2 which will be only trained with a VGG16 architecture for binary IQA of cytology images. The last layer was discarded. The results of model 2 are presented in the following table 2.

Table 2: Results of model 2 using VGG16 architecture to assess image quality of pap smear cells (cytology dataset).

	Accuracy (%)	**Precision (%)**	**recall (%)**	**AUC (%)**
VGG16	99.49 ± 0.72	99.50 ± 0.70	99.49 ± 0.72	99.96 ± 0.07

VGG16 in the multiclass quality assessment of blood cells images achieved the best performance in all metrics. This may have happened due to the relatively small number of images in our dataset used in this work. Thus, the high complexity of that the deeper networks (ResNet50 and MobileNet), may lead to overfit on train. The AUC metric is higher than accuracy, which may indicate that our classifier achieves good performance on the positive class (high AUC) at the cost of a high false negatives rate. The model proposed for classification of the pap smear images quality achieved a good performance for the binary image quality classification task with 99.49% of accuracy. This algorithm classifies the images according to the presence or absence of distortions, this way the classifier is focused on low-level notions of quality. The difference between the metrics of models 1 and 2 can be explained by the fact that in model 1 there are much more natural distortions in the images which increases the classification challenge, while our generated dataset for model 2 only contains 4 different types of distortions. To classify the images taking in account semantic complex concepts it is necessary to provide more information to the model.

5 Conclusions and Future Work

The use of IQA in biomedical applications is essential to help in optimization and improvement not only in image processing techniques but also on diagnostic algorithms. Nevertheless, the use of IQA methods in medical applications is very limited to low notions of quality, such as distortions or noise on the images. It is essential to collect more data and semantic information about image quality to build more robust and accurate algorithms to assess quality. Thus, new studies on cervical cancer using IQA techniques such as deep learning techniques should be encouraged due to the flexibility and capacity of these techniques and due to the literature gap about this subject. In this work it was demonstrated the outstanding performance of deep learning algorithms using CNN for NR-IQA in biomedical databases.

For future works, since a screening system is expected to be able to avoid misclassifying, artifacts must be added to the synthetic dataset of cytology to test the capacity of the CNN classifiers to detect that. In the future, it may be also interesting, add noise only in part of the image, and train with these examples. After that, analyze the activation maps and see if the explanation is consistent with the spatial placement of the noise.

Acknowledgements

This work is financed by the ERDF – European Regional Development Fund through the Operational Programme for Competitiveness and Internationalisation - COMPETE 2020 Programme and by National Funds through the Portuguese funding agency, FCT - Fundação para a Ciência e a Tecnologia within project “POCI-01-0145-FEDER-028857”.

References

[1] Ewert Bengtsson and Patrik Malm. Screening for cervical cancer using automated analysis of pap-smears. *Computational and mathematical methods in medicine*, 2014:842037, 03 2014.

[2] Zhi Lu, Gustavo Carneiro, Andrew Bradley, Daniela Ushizima, Masoud S. Nosrati, Andrea Bianchi, Cláudia Carneiro, and Ghassan Hamarneh. Evaluation of three algorithms for the segmentation of overlapping cervical cells. *IEEE Journal of Biomedical and Health Informatics*, 21:1–1, 01 2016.

[3] Teresa Conceição, Cristiana Braga, Luís Rosado, and Maria Vasconcelos. A review of computational methods for cervical cells segmentation and abnormality classification. *International Journal of Molecular Sciences*, 20:5114, 10 2019.