Diagnosis of Gastric Intestinal Metaplasia in Narrow-Band images with Fractal Bilinear Deep Learning Models

Maria Pedroso[1]
up201805037@edu.fc.up.pt

Miguel L. Martins[1]
miguel.l.martins@inesctec.pt

Diogo Libânio[2]
diogolibanio@med.up.pt

Mário Dinis-Ribeiro[2]
mdinisribeiro@gmail.com

Miguel Coimbra[1]
mcoimbra@fc.up.pt

Francesco Renna[1]
francesco.renna@fc.up.pt

[1] INESC-TEC and FCUP
University of Porto
Porto, Portugal

[2] CIDES/CINTESIS and FMUP
University of Porto
Porto, Portugal

Abstract

We propose two bilinear models to detect Gastric Intestinal Metaplasia (GIM) in narrow-band images that combine the embeddings of a pre-trained Deep Neural Network (DNN) with the outcome of a local texture descriptor based on fractal geometry. Our methods improve the DNN performance by a significant margin over several metrics (*e.g.*, area under the curve (AUC) 0.815 *vs*. 0.738) in a dataset comprised of EGD narrow-band images.

1 Introduction

Gastric cancer (GC) is the fifth most common type of cancer worldwide and is also the cause of the third highest number of cancer-related deaths. An early diagnosis of this cancer is crucial since it could potentially lead to a 40% reduction in mortality rates. Gastric Intestinal Metaplasia (GIM) is a critical precursor of GC that can be characterized during *Esophagogastroduodenoscopy* (EGD) by finding aberrant tissue in the gastric mucosa using Narrow-band Imaging (NBI) modality. Nevertheless, GIM detection is challenging since it relies on fine-grained details which results in low inter-observer concordance among clinicians. Thus, our goal was to find an automated tool uninfluenced by subjective factors, such as the Deep Neural Networks (DNN) that already obtained promising results in Upper Gastrointestinal (UGI) endoscopy-related problems. However, DNNs are dependent on high-quality data for training, which is not always available since it is expensive to collect. To address this problem, the importance of texture in GIM detection was considered. Thus, we propose two different approaches that consist of combining through a bilinear model a local texture description based on fractal geometry with the outcome of a DNN. The concept that the fusion of fractal descriptors and deep learning models might lead to a robust GIM detector was prompted by the valuable role of fractal dimension in identifying texture patterns in natural images [5] and by the discriminative features previously acquired in a closely related visual context specifically, in the characterization of polyps during colonoscopy [2].

2 Fractal Bilinear Deep Neural Network Models

2.1 Bilinear models

Following the definition in [4], a bilinear model consists in a quadruple $\mathcal{B} = (f_a, f_b, \mathcal{P}, \mathcal{C})$, where $f_a : \mathbb{R}^{H\times W\times C} \to \mathbb{R}^{k\times A}$ and $f_b : \mathbb{R}^{H\times W\times C} \to \mathbb{R}^{k\times B}$ are feature extracting functions, $\mathcal{P}$ is a pooling function and $\mathcal{C}$ is a classification function. The outcome of a feature extracting function is a feature vector which is obtained using an image and a given location. The core concept of this model is merging the results generated by these two feature functions using an outer product at every location within the image. The outer product is computed using $f_a(\mathbf{x})^{\mathrm{T}} f_b(\mathbf{x})$, where $\mathbf{x} \in \mathbb{R}^{H\times W\times C}$ and $(\cdot)^{\mathrm{T}}$ stands for the transpose operator. The pooling function is applied to aggregate the results, and for our experiences, the sum-pooling was considered. Since, it does not consider the location of the features, this function is *orderless* which is an important attribute for texture and fine-grained classification [4]. Finally, the classification function is applied to obtain the output of the bilinear model.

2.2 Multi-Fractal Spectrum

Since we are working with a finite resolution we can only compute an estimation of the fractal dimension. A usual way to compute this is assuming that the number of δ-covers M that span a fractal point set $\mathbb{F}$ vary proportionally to the power of δ, as $\delta \to 0$:

$$M(\mathbb{F},\delta) \approx k\delta^{-\beta} \implies \log M(\mathbb{F},\delta) \approx \log k - \beta \log \delta, \tag{1}$$

where $k, \beta \in \mathbb{R}$. Consequently, the *empirical estimation* of the fractal dimension β will be:

$$\beta = \lim_{\delta\to 0} \frac{\log M(\mathbb{F},\delta)}{-\log \delta}. \tag{2}$$

Regarding this, the computation of the fractal dimension can be done by determining the slope of the plot of $\log M(\mathbb{F},\delta)$ *versus* $\log\delta$, for an appropriate finite range of δ. Based on this dimension Yong Xu *et al.* [5] proposed the Multi-Fractal Spectrum (MFS) in order to obtain a more accurate and complete descriptor of $\mathbb{F}$.

Consider an image I defined over $\Omega \subseteq \mathbb{R}^2$ and let μ be a measure over Ω so that $\mu(\mathbf{x}, r) = kr^{\hat\beta(I,\mathbf{x})}$ for $\mathbf{x} \in \Omega$, where $\hat\beta(I,\mathbf{x}) \in \mathbb{R}$ is a density function and $k \in \mathbb{R}$. Then, the *local density function* or *Hölder exponent* is determined as:

$$\hat\beta(I,\mathbf{x}) = \lim_{r\to 0} \frac{\mu\big(B(I(\mathbf{x}),r)\big)}{\log r}, \tag{3}$$

where $B(I(\mathbf{x}),r)$ denotes a closed disk of length r around coordinate $\mathbf{x}$ in I. Then, we partition Ω and obtained the following sets:

$$\mathbb{F}_{\bar\beta} = \left\{\mathbf{x} \in \Omega : \hat\beta(I,\mathbf{x}) = \bar\beta\right\}. \tag{4}$$

The MFS is obtained by computing the fractal dimension using (2), for each categorization:

$$\mathrm{MFS}(I) = \left\{ \lim_{\delta\to 0} \frac{\log M(\mathbb{F}_{\bar\beta},\delta)}{-\log\delta} : \bar\beta \in \mathbb{R} \right\}. \tag{5}$$

In practice, we establish an appropriate range N and perform the computation of (5) by employing a uniform division of $[0,N]$ into m discrete bins that are evenly spaced. The definition used was the one proposed in [5] since it was demonstrated that it is invariant under the bi-Lipschitz map, which includes transformations very common in these images. Regarding the function μ, we also used the ones present in [5]:

- $\mu_1\big(B(I(\mathbf{x}),r)\big) = \int_{B(I(\mathbf{x}),r)} G_r * I(\mathbf{x})\, d\mathbf{x}$, where G_r is a Gaussian blur filter with variance r, and '$*$' is the convolution operator
- $\mu_2\big(B(I(\mathbf{x}),r)\big) = \int_{B(I(\mathbf{x}),r)} \sum_{i=1}^{4} g_i\big(G_r * I(\mathbf{x})\big)\, d\mathbf{x}$., where g_1, g_2, g_3, and g_4 are the differential operator for vertical, horizontal, diagonal, and anti-diagonal directions
- $\mu_3\big(B(I(\mathbf{x}),r)\big) = \int_{B(I(\mathbf{x}),r)} |\nabla^2\big(G_r * I(\mathbf{x})\big)| d\mathbf{x}$.

Finally, the MFS is the combination of μ_1, μ_2, and μ_3.

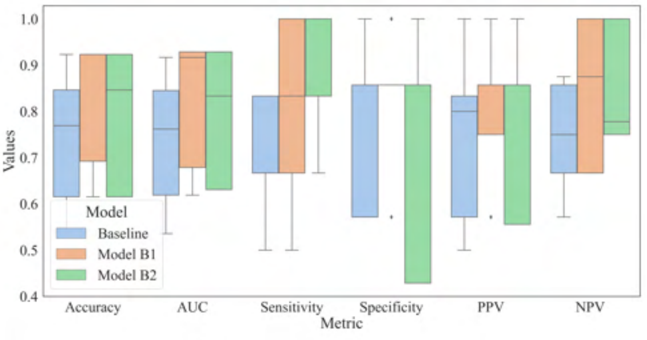

Figure 1: Boxplots with the metrics computed for each fold obtained in 5-fold cross-validation for the three different models.

2.3 Proposed approaches

For the two proposed models $\mathcal{B}_1$ and $\mathcal{B}_2$ the pooling function and the classification function were the same. The pooling function considered was the sum-pooling, while the classification function was defined as a Multi-Layer Perceptron (MLP) composed by two consecutive dense layers with rectified linear unit activation function followed by their respective dropout probabilities and an output layer which is a single neuron with a sigmoid activation function. In the two approaches the images were divided into patches to capture more local information.

The main difference between the two bilinear models $\mathcal{B}_1$ and $\mathcal{B}_2$ is the feature functions. On one hand, for $\mathcal{B}_1$, f_a is the output of the global average pooling of a pre-trained VGG-16 and f_b the maximum response over μ_1, μ_2, and μ_3. On the other hand, for $\mathcal{B}_2$, f_a is the embeddings of the VGG-16 and f_b the response over μ_1, μ_2, and μ_3 for all the patches. Thus, while $\mathcal{B}_1$ ensures that if there's a patch with a significant response, it signifies a segment of the image containing GIM, $\mathcal{B}_2$ preserves patch location information before computing the outer product allowing a clear spatial representation of pairwise interactions.

3 Experiments

3.1 Materials

For our experiences a dataset collected at the Gastroenterology department of Instituto Português de Oncologia, Porto (IPO-Porto) was used. Initially, the dataset was filtered concerning incorrect diagnosis of GIM, low resolution, and frames captured in WLI, and the final dataset was composed by a total of 125 high-quality NBI images, 65 classified as normal (- class) and 60 as GIM (+ class). Furthermore, a pre-processing procedure was applied to remove excessive black borders and delete the system status information. The dimension of the images was increased to $1078 \times 1351 \times 3$ using bilinear interpolation and then each image was divided into 7×7 non-overlapping grayscale patches with the shape of $154 \times 193 \times 1$, to be compatible with the shape of the embeddings of the VGG-16.

3.2 Experimental method

Besides the purposed approaches, following a previous work [3] we select a VGG-16 pre-trained in the ImageNet dataset was implemented as a baseline. In the three experiments, stratified 5-fold cross-validation was conducted, and for each cross-validation iteration, we generated the following partitions: 100 $(48+, 52-)$ samples for the train set, 12 $(6+, 6-)$ samples for the validation set and the 13 $(6+, 7-)$ remaining ones for the test set. In order to increase the amount of data we decided to use a simple data augmentation procedure on the training set that consisted in a random horizontal and vertical flips, and adding Principal Component Gaussian noise to the color channels [1]. The augmented training set resulted on a total average of 1214 samples $(582+, 632-)$. To estimate the MFS we set $r, \delta \in \{1,2,3,4,5,6,7,8\}$ and we defined $m = 26$ (number of bins to partition the interval $[0, N]$). All the MFS vectors obtained were normalized using the standardization method ($\mu = 0$ and $\sigma = 1$).

For evaluating the results 6 evaluation metrics were chosen: accuracy, Positive Predictive Value (PPV), Negative Predictive Value (NPV), Sensitivity, Specificity, and Area Under the Curve (AUC) (Fig.1).

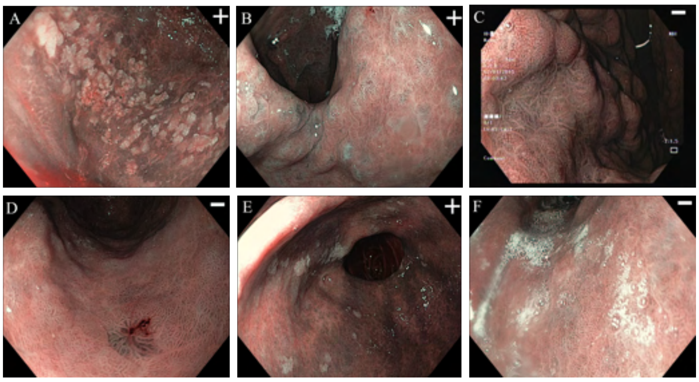

Figure 2: A, B, and C are images misclassified by the VGG-16, but correctly classified by $\mathcal{B}_1$ and $\mathcal{B}_2$; D; E, and F are the opposite. The true label is in the right top corner of the image.

4 Discussion

Concerning the results presented in Fig.1 we verified that there is a clear difference between the proposed approaches and the baseline. For all the metrics the proposed approaches exceed the baseline, except for the Specificity. Regarding $\mathcal{B}_1$ and $\mathcal{B}_2$ there is not a clear difference in performance since the values throughout each fold are very similar. In Fig.1 we also observed that the inter-fold variability is very significant because the interquartile range is large, except for the Specificity obtained with $\mathcal{B}_1$.

For a better understanding of the performance of the proposed approaches, we asses which images the baseline failed and our approaches not and the oppposite (Fig.2). As we noticed the baseline tends to fail in easy positive images (see image A and B in Fig.2) and it seems not able to capture the right texture pattern since it classified incorrectly images B and C from Fig.2. Concerning the proposed approaches we verified that they tend to misclassify normal images corrupted by the presence of noise (see image D and F in Fig. 2) and images in which a texture pattern is not clear (see image E in Fig. 2).

The work presented has essentially three limitations. Firstly, the choice of the patches and the level of overlap were chosen regarding the dimension of the embeddings of the VGG-16 in order to compute the outer product. Secondly, the images were labelled only for one expert. Finally, the dataset had no details about the patients which prevented us from doing a per-patient analysis.

5 Conclusions

We presented a novel DNN that incorporates explicit fractal descriptors to identify GIM within endoscopic imaging data. Regarding the higher metrics obtained, we conclude that our approaches represent a robust GIM detector even for a scarce dataset.

References

[1] Alex Krizhevsky et al. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25:1097–1105, 2012.

[2] Michael Häfner et al. Local fractal dimension based approaches for colonic polyp classification. *Medical Image Analysis*, 26(1):92–107, 2015. doi: 10.1016/j.media.2015.08.007.

[3] Miguel Martins et al. Diagnostic performance of deep learning models for gastric intestinal metaplasia detection in narrow-band images, 2023.

[4] Tsung-Yu Lin et al. Bilinear CNN models for fine-grained visual recognition. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 1449–1457, 2015. doi: 10.1109/ICCV.2015.170.

[5] Yong Xu et al. Viewpoint invariant texture description using fractal analysis. *International Journal of Computer Vision*, 83(1):85–100, 2009. doi: 10.1007/s11263-009-0220-6.