Classification of anatomical landmarks in the gastrointestinal tract with endoscopy images utilizing convolutional neural networks and triplet loss.

Lisle Faray de Paiva[1] **, Alan Carlos de Moura Lima**[1] **,**
Geraldo Braz Júnior[1] **, Anselmo Cardoso de Paiva**[1] **, Aristófanes Correa Silva**[1]

[1]Universidade Federal do Maranhão (UFMA)
Av. dos Portugueses, 1966 - Vila Bacanga, São Luís - MA, 65080-805

{lisle, alanlima, geraldo, paiva, ari}@nca.ufma.br

Abstract. *According to the World Health Organization, 8 million deaths are counted due to gastrointestinal diseases annually. Automatic detection of gastrointestinal landmarks is a task that can further help medical professionals reducing cost and time in exploratory exams. Computer-aided detection and diagnosis systems have been widely explored in the scientific field. However, it takes a lot of processing power to achieve satisfactory results. In order to overcome this problem, this work uses a simple Convolutional Neural Network together with the Triplet Loss cost function to extract image characteristics of 3 gastrointestinal anatomical landmarks (z-line, pylorus, and cecum) to classify those images.For the training it's used the dataset Kvasir-v2, obtaining 96,60% of Precision, 97,71% of Accuracy, 96,91% of Specificity, 98,61% of Recall 97,59% of F1-score.*

Resumo. *De acordo com a Organização Mundial de Saúde, anualmente são contabilizadas 8 milhões de mortes devido a doenças do trato gastrointestinal. A detecção automática das marcações anatômicas é uma tarefa que pode auxiliar profissionais da área da saúde, reduzindo custo e tempo em exames exploratórios. Sistemas de detecção e diagnóstico auxiliados por computador têm sido vastamente explorado no âmbito científico. No entanto é necessário muito poder de processamento para atingir resultados satisfatórios. Com o intuito de contornar esse problema, este trabalho utiliza uma Rede Neural Convolucional simples em conjunto da função de custo Triplet Loss para extrair características de imagens de 3 marcações anatômicas gastrointestinais (z-line, pylorus e cecum) para classificação dessas imagens. Para o treinamento é utilizada a base de dados Kvasir-v2, obtendo 96,60% de Precisão, 97,71% de Acurácia, 96,91% de Especificidade, 98,61% de Sensibilidade e um F1-score de 97,59%.*

1. Introdução

O trato gastrointestinal inclui todos os órgãos do sistema digestivo. Trata-se da via responsável por todo o processo digestivo, sendo formado pela boca, esôfago, estômago, intestino delgado, intestino grosso e ânus. Diversas doenças podem afetar essa importante via, como por exemplo: câncer, hemorragias, esofagite, pólipos e colite ulcerosa.

Doenças do trato gastrointestinal são prevalentes em todo o mundo, causando uma alta mortalidade, requerendo a utilização de cuidados específicos dos sistemas de

saúde. Em todo o mundo, no ano de 2015, houve aproximadamente 8 milhões de mortes relacionadas a doenças no trato gastrointestinal [Chan et al. 2019]. De acordo com [Organizações das Nações Unidas 2018], no Brasil há cerca de 19% da população acometida com pelo menos um tipo de doença relativa ao trato gastrointestinal independentemente do gênero sexual.

A endoscopia gástrica é o exame comumente utilizado para localização e prevenção de doenças do trato gastrointestinal [Deeba et al. 2020]. De acordo com [Wittenberg et al. 2019] a endoscopia é realizada por um profissional da saúde através de um endoscópico (tubo flexível munido de uma pequena câmera), onde filma-se a mucosa interna do trato digestivo a partir da boca (endoscopia gástrica) ou do reto (colonoscopia). Durante o exame, é necessário cuidado para identificar os órgãos e as possíveis patologias, e para isso os especialistas utilizam as marcações anatômicas, essenciais por servirem como ponto de referência e por descreverem a localização de uma determinada região [Cogan et al. 2019].

Algumas das principais marcações anatômicas do trato digestivo são *z-line*, *pylorus* e *cecum*. A *z-line* indica a junção esofagogástrica entre a mucosa escamosa do esôfago e a mucosa do estômago; o *pylorus* conecta o estômago ao intestino delgado. Já o *cecum* representa, para a colonoscopia (endoscopia gástrica começando pelo reto), a região final do exame, o fim do intestino grosso [Gamage et al. 2019]. Essas marcações também podem ser regiões típicas de presença de patologias como úlceras ou inflamações [Pogorelov et al. 2017].

O reconhecimento das marcações anatômicas é vital para o exame de endoscopia dada a necessidade de se localizar os órgãos e estruturas internas do trato gastrointestinal, uma vez que elas funcionam como ponto de referência [Cogan et al. 2019]. O reconhecimento é feito através do especialista conduzindo o exame. Com o objetivo de auxiliar o exame de endoscopia, pode-se utilizar técnicas computacionais para detecção das marcações anatômicas no trato gastrointestinal como este trabalho propõe.

Diversas técnicas computacionais já foram desenvolvidas com o uso de aprendizagem de máquina com o intuito de analisar imagens de endoscopia para classificar ou detectar estruturas internas e patologias, como em: [Itoh et al. 2018, de Lange et al. 2018, Alaskar et al. 2019]. Em [Takiyama et al. 2018] foi realizada a classificação de marcações anatômicas de regiões como a laringe, esôfago, estômago e duodeno, alcançando resultados para as imagens de laringe e esôfago de 1.00 em AUC; e para as imagens de estômago e duodeno de 0.99 em AUC. Já, em [Cogan et al. 2019] foi realizada a classificação das imagens das marcações anatômicas da base kvasir-v2, utilizando as redes neurais convolucionais: Inception-v4, Inception-ResNet-v2 e NASNet, alcançando acurácias respectivas de 0.9845, 0.9848 e 0.9735.

Redes como a Inception ou ResNet previamente citadas são robustas e requerem muito tempo e recursos para atingirem resultados satisfatórios. Tentando contornar esse problema este trabalho propõe analisar o efeito do uso de uma Rede Neural Convolucional (CNN, do inglês *Convolutional Neural Network*) com uso da função de perda triplet [Weinberger and Saul 2009] em diferentes versões de imagens endoscópicas contendo marcações anatômicas (*z-line*, *pylorus* e *cecum*) para a classificação dessas regiões anatômicas. Nos experimentos realizados foi utilizada a base de imagens Kvasir-v2 [1]

[Pogorelov et al. 2017].

Esse artigo está organizado da seguinte forma: a seção 2 apresenta a metodologia proposta para o desenvolvimento deste trabalho, informando as técnicas utilizadas no pré-processamento das imagens de endoscopia e como foi realizada a construção do classificador baseado em uma rede neural convolucional. A seção 3 mostra os resultados alcançados após a aplicação da metodologia proposta e a seção 4 apresenta as conclusões e trabalhos futuros.

2. Metodologia

Este trabalho apresenta uma metodologia para classificação das marcações anatômicas do trato gastro intestinal como apresentada na Figura 1. Primeiramente é realizada a etapa de pré-processamento das imagens. São geradas 4 versões da base de dados: em escala de cinza, em escala de cinza em conjunto com a equalização adaptativa de histograma e em escala de cinza em conjunto com a equalização de histograma e colorida adotando o padrão RGB. Cada versão gerada é então alimentada a uma rede neural convolucional que treina com o uso da função de perda *triplet*. Após o treinamento, são extraídos os *embeddings* para classificação através do classificador *Random Forest*.

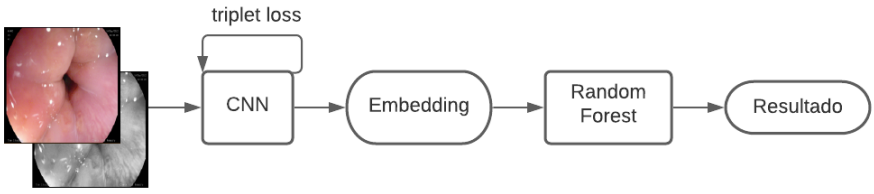

Figura 1. Fluxograma da metodologia proposta.

2.1. Aquisição das imagens

A base de dados utilizada nessa metodologia é a Kvasir-v2 [Pogorelov et al. 2017]. Ela é dividida em 8 classes, com 1000 imagens para cada classe. No entanto, para este trabalho, são utilizadas somente as três 3 classes relacionadas às marcações anatômicas: *pylorus*, *z-line* e *cecum*, exemplificadas na Figura 2, totalizando 3000 imagens com 1000 imagens por marcação anatômica. As imagens foram distribuídas para o treinamento da CNN com 90% para a fase do treino e validação e 10% para testar o modelo.

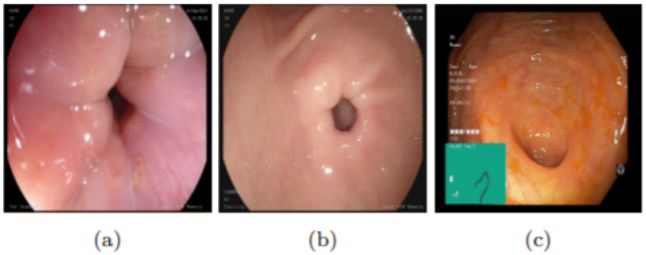

Figura 2. Exemplos das marcações anatômicas: (a) *z-line* (b) *pylorus* (c) *cecum*.

[1]https://datasets.simula.no/kvasir/

2.2. Técnicas de melhoramento da qualidade das imagens

Com o intuito de melhorar a qualidade das imagens em escala de cinza, são utilizadas 2 técnicas de pré-processamento separadamente: equalização de histograma e o CLAHE (*Contrast Limited Adaptaive Histogram Equalization*).

A equalização de histograma é uma técnica de transformação de intensidade de pixeis com o objetivo de balancear os níveis de cinza em uma imagem [Abdullah-Al-Wadud et al. 2007]. O CLAHE é uma variação de equalização de histograma adaptativo em que a amplificação do contraste é limitada uma vez que em técnicas de equalização de histograma adaptativo comuns acabam por amplificar ruídos em imagens [Pizer et al. 1987]. Ambas técnicas são recomendadas para imagens em escala de cinza, portanto são utilizadas para verificar qual possui melhor desempenho.

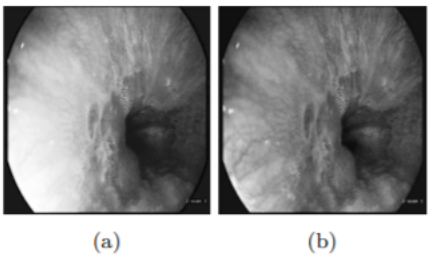

Figura 3. Imagens após a realização das técnicas de pré-processamento: (a) Equalização de Histograma (b) CLAHE.

Ambas técnicas possuem seus pontos fortes e fracos. Como observado na Figura 3, a equalização de histograma acaba por destacar mais as áreas brancas das escuras, isso se dá devido as bordas pretas das imagens utilizadas. Já com o CLAHE é possível observar que, por conta do seu limite de contraste, ele gera uma imagem mais uniformizada, realçando o sinal da imagem sem amplificar exorbitantemente possíveis ruídos.

Além das imagens em escalas de cinza, são utilizadas imagens coloridas utilizando o padrão de cores RGB com o intuito de analisar se há perda de informação com a perda de cor. Para essas não foi necessário aplicar nenhuma técnica de melhoramento.

2.3. Função *triplet loss*

A *Triplet Loss* é uma função de perda que otimiza um espaço euclidiano de forma que os dados com a mesma identidade sejam mais próximos uns dos outros do que aqueles com diferentes identidades. A ideia por trás da *Triplet Loss* é maximizar a distância entre os dados com identidades diferentes e minimizar a distância entre dados com a mesma identidade dentro do espaço vetorial.

Ela é formada por um terceto $Triplet = (\hat{A}ncora, Positivo, Negativo)$, onde as imagens $\hat{A}ncora$ e $Positivo$ são da mesma classe e a imagem $Negativo$ é de uma classe distinta. No começo do treinamento é calculada a distância das imagens $\hat{A}ncora, Positivo$ e $Negativo$ em um espaço vetorial. O intuito da *Triplet Loss* é de minimizar a distância entre a $\hat{A}ncora$ e o $Positivo$, representando imagens de mesma classe como próximas e maximizar a distância entre a $\hat{A}ncora$ e o $Negativo$, representando classes diferentes como distantes como demonstrado na Figura 4. Como descrito em [Schroff et al. 2015], a função que está sendo minimizada neste artigo está especificado na Equação 1:

$$L = \sum_{i}^{N} [\|f(x_i^a) - f(x_i^p)\|_2^2 - \|f(x_i^a) - f(x_i^n)\|_2^2 + \alpha]_+, \quad (1)$$

onde: x_i^a é a imagem âncora, x_i^p é a imagem positiva, x_i^n é a imagem negativa, $f(x) \in R^d$ equivale ao *embedding* incorporando uma imagem x em um espaço euclidiano d-dimensional e α representa a margem entre os pares positivos e negativos.

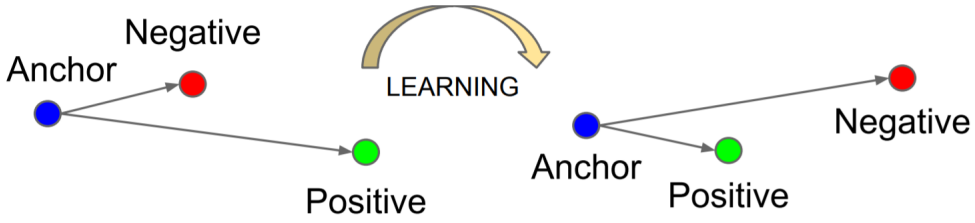

Figura 4. Função de custo Triplet Loss.

2.4. Classificação das imagens

Após o pré-processamento, as imagens são enviadas para uma CNN base cuja arquitetura pode ser observada na Figura 5, para a etapa de treinamento com o uso da função de custo *Triplet Loss*. Com a conclusão do treinamento, são extraídos *embeddings* para a classificação utilizando o classificador *Random Forest* [Pedregosa et al. 2011]. Os *embeddings* gerados são espaços euclidianos onde é possível melhor discernir as classes entre si.

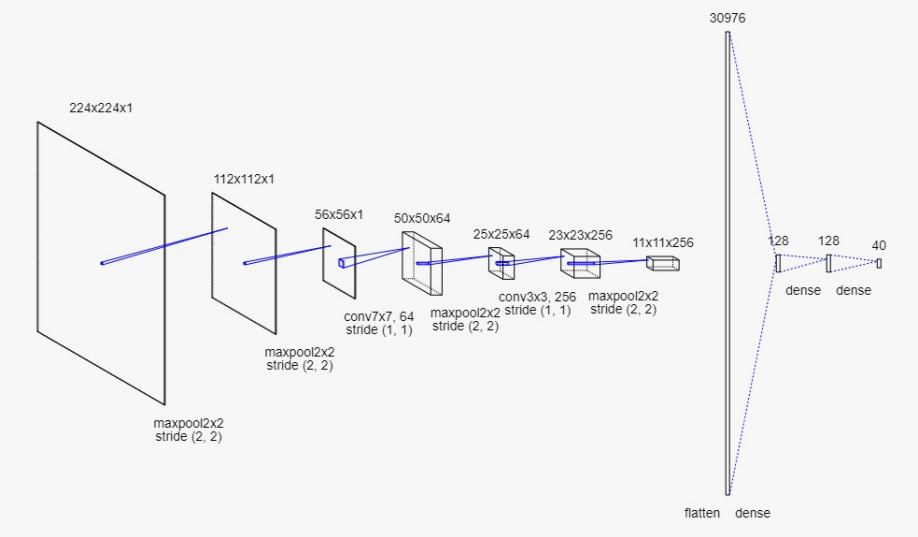

Figura 5. Arquitetura da CNN proposta.

3. Resultados

Para avaliar a metodologia proposta foram usadas as seguintes métricas de avaliação: acurácia, sensibilidade, especificidade, precisão e *F1-score*, descritas nas Equações 2, 3, 4, 5 e 6, respectivamente. Assim, busca-se uma metodologia que tenha a capacidade de

classificar corretamente as três classes (sensibilidade) e ao mesmo tempo distinguir as classes entre si (especificidade).

$$ACU = \frac{TP + TN}{TP + TN + FP + FN} \tag{2}$$

$$SEN = \frac{TP}{TP + FN} \tag{3}$$

$$ESP = \frac{TN}{TN + FP} \tag{4}$$

$$PRE = \frac{TP}{TP + FP} \tag{5}$$

$$F1 = \frac{2 \cdot PRE \cdot SEN}{PRE + SEN} \tag{6}$$

Analisando a Tabela 1 pode-se observar que o experimento utilizando as imagens em RGB obteve resultados superiores aos obtidos com o uso das imagens em escala de cinza. Fazendo um comparativo entre os resultados das imagens em escala de cinza, pode-se observar que para as imagens que sofreram a equalização de histograma houve uma melhora na especificidade, ou seja, houve uma redução das imagens classificadas para classes as quais não pertencem, todavia ele foi o experimento que obteve a sensibilidade com menor valor.

Relacionado as imagens com o processamento utilizando a técnica CLAHE, pode-se observar que embora seja uma técnica que tende a melhorar a qualidade do contraste das imagens de forma adaptativa, não foi suficiente para resultar em uma melhora das métricas em relação às outras técnicas aqui desenvolvidas. Das imagens em escala de cinza, a sem técnicas de pré-processamento foi a que obteve melhores métricas, não obstante ao compará-la com os resultados obtidos no experimento com imagens em RGB, pode-se observar que ocorreu uma perda significativa de informação ao converter as imagens coloridas para tons de cinza.

Tabela 1. Comparação dos resultados obtidos com o método proposto e o trabalho relacionado para o problema de classificação de marcações anatômicas em imagens de endoscopia. Os valores em negrito se referem aos melhores resultados encontrados naquela métrica.

Tests	Sensibility	Specificity	Precision	Accuracy	F1-score
GrayScale	87.50%	91.36%	90.00%	89.54%	88.73%
Histogram Equalization	83.33%	93.17%	91.60%	88.52%	87.27%
CLAHE	84.03%	90.12%	87.94%	97.91%	87.02%
RGB	**98.61**%	96.91%	**96.60**%	97.71%	**97.59**%
Cogan et al. 2019	93.9%	**99.1**%	93.8%	**98.45**%	92.9%

Em comparação com o estado da arte [Cogan et al. 2019], um trabalho sobre a detecção das mesmas marcações anatômicas utilizando a Inception-ResNet-v2, Inception-v4 e a NASNet em imagens coloridas do trato gastrointestinal, os resultados obtidos em imagens coloridas como na Precisão, Sensibilidade e F1-score, ultrapassam as métricas de [Cogan et al. 2019], mostrando que com o uso da função de custo *Triplet*

Loss pode-se obter resultados satisfatórios sem a necessidade de uma rede neural mais profunda e um grande poder computacional. No entanto ainda há espaço para melhoria, visto que a Acurácia e Especificidade não ultrapassam os valores encontrados em [Cogan et al. 2019].

4. Conclusão

O objetivo principal deste trabalho foi utilizar uma CNN com a função triplet loss para extração de características para classificar 3 grupos de imagens de endoscopia referentes às marcações anatômicas do trato gastrointestinal. Nos experimentos realizados as características foram classificadas com um algoritmo Random Forest. Resultados satisfatórios foram alcançados com a utilização de imagens em escala de cinza e RGB, além de apresentado um estudo sobre a utilização e influência de técnicas de pré-processamento.

Em trabalhos futuros pretende-se aplicar a imagens coloridas a técnica de aumento de dados a fim de gerar novas amostras para amplificar o treinamento do nosso modelo. Além disso, será analisada a utilização de uma CNN mais profunda, em substituição à MLP como por exemplo: VGG16, ResNet, EfficientNet.

Pretende-se também investigar se outros classificadores conseguem obter resultados comparáveis ao Random Forest e explorar a utilização de uma metodologia completamente baseada em arquiteturas de Deep Learning (sem extração de características e uso de classificador externo).

Referências

Abdullah-Al-Wadud, M., Kabir, M. H., Dewan, M. A. A., and Chae, O. (2007). A dynamic histogram equalization for image contrast enhancement. *IEEE Transactions on Consumer Electronics*, 53(2):593–600.

Alaskar, H., Hussain, A., Al-Aseem, N., Liatsis, P., and Al-Jumeily, D. (2019). Application of convolutional neural networks for automated ulcer detection in wireless capsule endoscopy images. *Sensors*, 19(6).

Chan, J. S. H., Chao, A. C. W., Cheung, V. C. H., Wong, S. S. K., Tang, W., Wu, J. C. Y., Chan, H. L. Y., Chan, F. K. L., Sung, J. J. Y., and Ng, S. C. (2019). Gastrointestinal disease burden and mortality: A public hospital-based study from 2005 to 2014. *Journal of gastroenterology and hepatology*, 34(1):124–131.

Cogan, T., Cogan, M., and Tamil, L. (2019). Mapgi: Accurate identification of anatomical landmarks and diseased tissue in gastrointestinal tract using deep learning. *Computers in Biology and Medicine*, 111:103351.

de Lange, T., Halvorsen, P., and Riegler, M. (2018). Methodology to develop machine learning algorithms to improve performance in gastrointestinal endoscopy. *World Journal of Gastroenterology*, 24(45):5057–5062.

Deeba, F., Bui, F. M., and Wahid, K. A. (2020). Computer-aided polyp detection based on image enhancement and saliency-based selection. *Biomedical Signal Processing and Control*, 55:101530.

Gamage, C., Wijesinghe, I., Chitraranjan, C., and Perera, I. (2019). Gi-net: Anomalies classification in gastrointestinal tract through endoscopic imagery with deep learning. In *2019 Moratuwa Engineering Research Conference (MERCon)*, pages 66–71. IEEE.

Itoh, T., Kawahira, H., Nakashima, H., and Yata, N. (2018). Deep learning analyzes helicobacter pylori infection by upper gastrointestinal endoscopy images. *Endoscopy International Open*, 06(02):E139–E144.

Organizações das Nações Unidas (2018). Colorectal cancer. international agency for research on cancer.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Pizer, S. M., Amburn, E. P., Austin, J. D., Cromartie, R., Geselowitz, A., Greer, T., ter Haar Romeny, B., Zimmerman, J. B., and Zuiderveld, K. (1987). Adaptive histogram equalization and its variations. *Computer vision, graphics, and image processing*, 39(3):355–368.

Pogorelov, K., Randel, K. R., Griwodz, C., Eskeland, S. L., de Lange, T., Johansen, D., Spampinato, C., Dang-Nguyen, D.-T., Lux, M., Schmidt, P. T., Riegler, M., and Halvorsen, P. (2017). Kvasir: A multi-class image dataset for computer aided gastrointestinal disease detection. MMSys'17, page 164–169, New York, NY, USA. Association for Computing Machinery.

Schroff, F., Kalenichenko, D., and Philbin, J. (2015). Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823.

Takiyama, H., Ozawa, T., Ishihara, S., Fujishiro, M., Shichijo, S., Nomura, S., Miura, M., and Tada, T. (2018). Automatic anatomical classification of esophagogastroduodenoscopy images using deep convolutional neural networks. *Scientific Reports*, 8(1).

Weinberger, K. Q. and Saul, L. K. (2009). Distance metric learning for large margin nearest neighbor classification. *Journal of machine learning research*, 10(2).

Wittenberg, T., Zobel, P., Rathke, M., and Mühldorfer, S. (2019). Computer aided detection of polyps in whitelight-colonoscopy images using deep neural networks. *Current Directions in Biomedical Engineering*, 5(1):231–234.