IDENTIFICAÇÃO DE EDIFICAÇÕES A PARTIR DE IMAGENS DE SENSORES REMOTOS UTILIZANDO A ARQUITETURA COMPUTACIONAL MASK R-CNN

André Estevam Costa Oliveira [1] e Wagner Santos de Almeida [2]

[1]Universidade de Brasília, Brasília-DF, andre.estevam.unb@gmail.com; [2]Universidade de Brasília, Brasília-DF, wagnersa_98@yahoo.com

RESUMO

Diante da evolução da coleta de dados por sensores remotos e do poder de processamento dos computadores, o campo do Aprendizado de Máquina está em rápida evolução devido à ampla aplicabilidade de suas tecnologias. No campo do sensoriamento remoto, tais tecnologias manifestam-se principalmente em tarefas de classificação temática e segmentação, sobretudo na extração de *features* e classificadores. Tais avanços são importantes para a agricultura de precisão, planejamento urbano, monitoramento inteligente, entre outras. O objetivo deste trabalho é o emprego de tecnologias emergentes da área de Aprendizado de Máquina, especificamente a arquitetura estado-da-arte "Mask R-CNN", para a identificação e segmentação automatizada de edificações a partir de imagens de sensores remotos, com o intuito de detectar edificações urbanas e outras intervenções antrópicas em mapas de uso e cobertura do solo.

Palavras-chave — Sensoriamento Remoto, Geoprocessamento, Aprendizado de Máquina, Mask R-CNN, CBERS-4A.

ABSTRACT

Given the rapid evolution of data collection and computer processing power, the field of Machine Learning is on the rise given its technologies' broad applicability, permeating many different areas of knowledge. In the field of Remote Sensing, such technologies manifest mainly through thematic classification and segmentation tasks, specifically feature extraction and classifiers. These advances are important for precision agriculture, urban planning, efficient monitoring etc. The objective of this article is the deployment of emerging technologies, specifically the state-of-the-art architecture "Mask R-CNN", for automated identification and segmentation of edifications in remote sensing images, with the intent of detecting edification and other anthropologic interventions in land-cover and land-use maps.

Key words *— Remote Sensing, Geoprocessing, Machine Learning, Mask R-CNN, CBERS-4A.*

1. INTRODUÇÃO

Perante a expansão da indústria aeroespacial, imagens ópticas de média e alta resolução espacial estão cada vez mais acessíveis em um amplo espectro espacial e temporal. Paralelamente, a rápida evolução de transmissores de longo alcance, dispositivos de localização e sensores de imageamento tornam acessível economicamente e logisticamente a utilização de Veículos Aéreos Não Tripulados (VANT) [1]. Tais fatores, entre outros, permitem uma crescente disponibilidade de grandes volumes de dados de Sensoriamento Remoto (SR), os quais estimulam o desenvolvimento e aplicação de técnicas robustas de análise de imagens capazes de abordar eficaz e eficientemente tarefas cada vez mais complexas como reconstrução 3D [2], super-resolução [3] e, particularmente, detecção de objetos [4]. O campo o qual vem demonstrando excelência em algumas destas tarefas e tem atraído interesse de diversos setores da sociedade e da academia é o Aprendizado de Máquina e uma de suas áreas correlacionadas, a Visão Computacional, a qual dedica-se à estruturação de sistemas visuais artificiais capazes de interpretar e compor entendimentos complexos a respeito do mundo real [5].

No que tange o ordenamento territorial, agentes públicos encarregados da fiscalização da expansão de assentamentos populacionais, ou da própria confecção de mapas cartográficos de uso e cobertura do solo, podem beneficiar-se de sistemas automatizados de análise de imagens para uma resposta rápida e eficaz. Portanto, o desenvolvimento de um modelo de detecção de ocupações antrópicas em imagens pode ser de grande valia para a prevenção da grilagem de terra, gerenciamento do solo e o desenvolvimento urbano sustentável de modo geral.

A tarefa de detecção de objetos (*Object Detection* - OD) consiste na localização e classificação de distintos alvos em uma imagem e é endereçada por diversas metodologias permeando diversos campos [4]. No campo de Aprendizado de Máquina (AM), entre as metodologias consideradas como estado da arte, pode-se divisar abordagens de métodos de uma ou duas etapas, esta última baseada na proposição de regiões (*region-proposals*), que podem apresentar maior precisão na detecção de objetos e, em contrapartida, maior tempo de execução [4]. Apesar de tais métodos apresentarem ótimos resultados em imagens naturais, sua aplicabilidade em imagens de SR ainda não é amplamente explorada [6].

Diante do exposto, faz-se pertinente avaliar a abordagem em duas etapas para detecção de objetos em imagens de SR de alta e média resolução espacial, avaliando sua utilização em diferentes cenários de disponibilidade de dados, visando a aplicação de tais tecnologias no que tange o ordenamento territorial.

Este trabalho objetiva desenvolver dois modelos baseados na arquitetura "Mask R-CNN" proposto por [7], para a detecção e segmentação de ocupações antrópicas, avaliando sua performance em imagens de sensores orbitais e em fotografias aéreas de alta resolução espacial, com o intuito final de auxiliar a confecção de mapas de uso e cobertura do solo e o monitoramento do gerenciamento do solo.

2. MATERIAL E MÉTODOS

Para o desenvolvimento deste trabalho fora utilizadas imagens do satélite CBERS-4A, de 8 m de resolução espacial para as bandas multiespectrais e de 2 m para a banda pancromática, abrangendo o Distrito Federal, capturadas em julho de 2020 e disponibilizadas pelo INPE. Os dados vetoriais georreferenciadas de edificações do Distrito Federal foram cedidos pela Empresa Distrital Terracap, nas quais selecionou-se algumas regiões do Distrito Federal para a fase de extração de amostras conforme a densidade de edificações e disponibilidade de dados georreferenciados.

Adicionalmente se utilizou do *dataset* estruturado por [6], contendo amostras de alta qualidade de várias regiões urbanizadas, além de imagens aéreas de 0,24 m de resolução espacial abrangendo a região do Plano Piloto, cedidas pela Terracap e obtidas em meados de 2018 e 2019, somado aos dados vetoriais já mencionados.

Para o pré-processamento dos dados raster e correção dos dados vetoriais (devido a discrepâncias nas datas de captura das imagens e constituição das anotações), os ambientes computacionais dos sistemas de informações geográficas (SIG) Spring e ArcGIS foram utilizados.

No processamento e formatação dos dados, desenvolvimento e implementação do modelo, foi utilizado a linguagem de programação Python somada ao *framework* PyTorch, um pacote de códigos especializados em algoritmos de AM, utilizando a ferramenta Google Colab como ambiente de execução, a qual disponibiliza uma GPU (*Graphics Processing Unit*), visto a exigência de um alto poder de processamento.

2.1. Banco de dados

As imagens do CBERS passaram por processo de fusão, combinando uma composição de bandas RGB com a pancromática, de maior resolução espacial, a fim de se obter uma imagem de bandas RGB de resolução espacial igual ao da pancromática, neste caso, de 2 metros. Paralelamente, os dados vetoriais das edificações passaram por processo de rasterização, mantendo a resolução da matriz resultante a mesma do raster produto da fusão, ou seja, 2 metros.

A partir daí a constituição das amostras se deu através do ArcGIS, através do qual foi possível a extração sistemática de fragmentos do raster produto da fusão e correspondente fragmento do raster produto da rasterização.

Tal procedimento se deu da mesma forma para ambos os *datasets* gerados (de média e alta resolução espacial), variando apenas o raster de bandas RGB. Para as imagens orbitais decidiu-se por extrair amostras de dimensão 256x256 pixels, pois para maiores dimensões a quantidade de amostras seria muito limitada e frequentemente abrangendo regiões de altíssima densidade de edificações. Para as imagens aéreas decidiu-se por extrair amostras de dimensões 300x300 pixels, de forma a garantir quantidade satisfatória de amostras. Ao todo obteve-se 483 amostras das imagens orbitais: 408 para treinamento, 34 para validação, 36 para teste e 4 descartadas devido à ausência de instâncias. Foram utilizadas 11058 amostras das imagens aéreas: 9952 para treinamento, 547 para validação, 547 para teste e 12 descartadas devido à ausência de instâncias. Em ambos os conjuntos as amostras passaram por processo de normalização. A Figura 1 exemplifica uma amostra originada a partir das imagens CBERS e a Figura 2 das imagens aéreas.

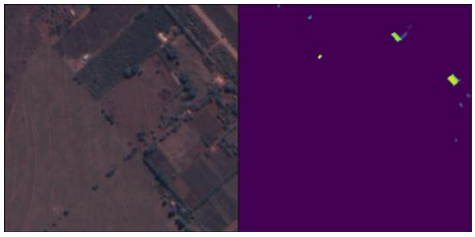

Figura 1. Amostra extraída de imagens do satélite CBERS-4A.



Figura 2. Amostra extraída de imagens aéreas cedidas pela Terracap.

2.2. Modelo

A arquitetura "Mask R-CNN", na qual os modelos deste trabalho se espelharam, constitui-se em uma série lógica de algoritmos do Aprendizado de Máquina, cada qual responsável por determinada operação. As Redes Neurais Artificiais (RNA's), Redes Neurais Convolucionais (CNN), *Region Proposal Network* (RPN) entre outros algoritmos compõem a estrutura base da arquitetura "Mask R-CNN" [7].

A arquitetura *backbone* utilizada, encarregada da etapa de extração de *features*, foi a ResNet-50. Dotada de uma Rede Profunda de Camadas Convolucionais (*Deep Convolutional*

Neural Network - DCNN), a arquitetura em questão permite extração de informações em um alto nível de abstração [8].

A *Region Proposal Network* (RPN) foi introduzida na arquitetura por [9] como uma solução para a antes custosa etapa de proposição de RoI's, além de integrar a etapa no processo de treinamento do modelo. Esta etapa é especialmente sensível em relação ao tamanho das instâncias pois, conforme é possível observar na Figura 3, a região da imagem original que cada pixel do *feature map* representa (*receptive field*) varia enormemente de acordo com a resolução espacial, fazendo do estabelecimento de tamanhos ótimos para as âncoras vital para um bom desempenho do modelo [6]. Para o modelo avaliado em imagens de média resolução espacial foram definidos os tamanhos: 2, 4, 8, 16 e 32, com escalas de 0,25, 0,5 e 1; e, para o avaliado em imagens de alta resolução: 8, 16, 32, 64 e 128 pixels com escalas de 1, 1,5 e 2, somando 15 âncoras por pixel do *feature map*.

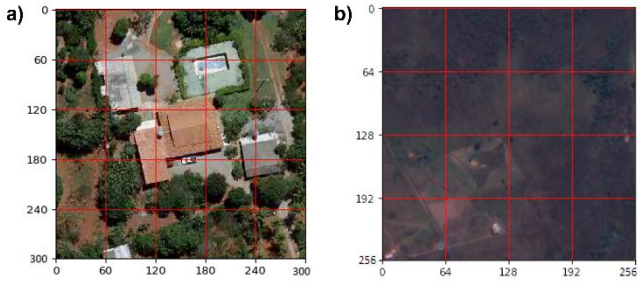

Figura 3. Ilustração do *receptive field* para uma imagem de alta resolução espacial (a) e para uma de média resolução (b).

Os demais componentes, como a rede completamente conectada (*Fully Connected* - FC) e a *Fully Convolutional Network* (FCN), foram implementadas conforme visto no trabalho original [7], portanto sem quaisquer alterações em relação a quantidade de camadas ou outros hiperparâmetros.

O treinamento do modelo é supervisionado. Como algoritmo de otimização foi o utilizado o Gradiente Descendente Estocástico (GDE). O peso inicial utilizado foi aquele treinado no *dataset* ImageNet, disponível abertamente. O modelo avaliado em imagens de alta resolução espacial passou por treinamento inicial com o *dataset* estruturado por Mohanty *et al.* [6], utilizando da estratégia *transfer learning*. Não foi possível a aplicação da mesma estratégia para o modelo avaliado em imagens do CBERS-4A pois a resolução espacial das amostras do *dataset* de Mohanty *et al.* [6] é de 0,3 m, sendo, portanto, muito discrepante da resolução das amostras do CBERS-4A. A quantidade de épocas, Taxa de Treinamento (*Learning Rate* - Lr), camadas treinadas entre outros hiperparâmetros passaram por extensivo processo de experimentação.

Em diversas aplicações o treinamento das camadas finais é o suficiente para o objetivo da aplicação, pois os pesos iniciais são geralmente treinados em imagens naturais, portanto seria necessário apenas que o modelo "aprenda" as relações de *features* mais abstratas localizadas mais profundamente na rede. No entanto, para casos mais específicos como imagens de SR, é importante o treinamento de camadas mais profundas devido à discrepância entre as características básicas de cada tipo de imagem [3]. Diante desta necessidade, o modelo avaliado em imagens de alta resolução espacial passou por poucas iterações devido às limitações no tempo de execução do ambiente Google Colab, o qual restringe o usuário a aproximadamente 36 horas de tempo no uso da GPU.

3. RESULTADOS E DISCUSSÃO

O modelo avaliado em imagens orbitais sofreu sobreajuste, isto é, o modelo não foi capaz de generalizar além das amostras que este foi exposto [5]. As seguintes estratégias foram utilizadas na tentativa de o treinamento convergir, além da regularização: modificação de hiperparâmetros (taxa de treinamento, camadas treinadas, número de épocas e tamanho das âncoras), diminuição da taxa de subamostragem através das camadas *pooling*, redimensionamento das amostras (interpolação bilinear) e técnicas de aumento de dados. O gráfico contido na Figura 4 mostra o comportamento do erro, ou seja, distância entre o inferido e o real, através do treinamento e validação, o qual se repete mesmo após aplicação das estratégias mencionadas.

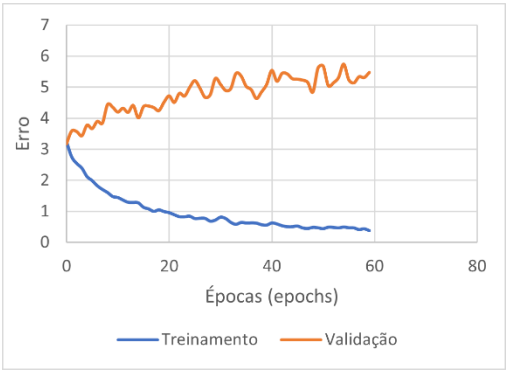

Figura 4. Comportamento do erro através do treinamento e validação do modelo utilizando imagens de média resolução espacial.

A principal causa do sobreajuste pode ser devido à resolução espacial da imagem, pela qual as instâncias de edificação possuíam tamanhos muito pequenos, ocupando, em sua maioria, menos de 20 pixels [3], fornecendo informação insuficiente para a RPN e algoritmos adiante. Outros fatores como quantidade reduzida de amostras e anotações incorretas ou faltantes podem ter contribuído para o comportamento observado.

Já o modelo avaliado em imagens de alta resolução espacial apresentou convergência durante todo o processo de treinamento. Utilizando a métrica IoU entre 0,5 e 0,95, seguindo os padrões do *Common Objects in Context* (COCO), obteve-se os valores de 0,536 e 0,612 para Precisão

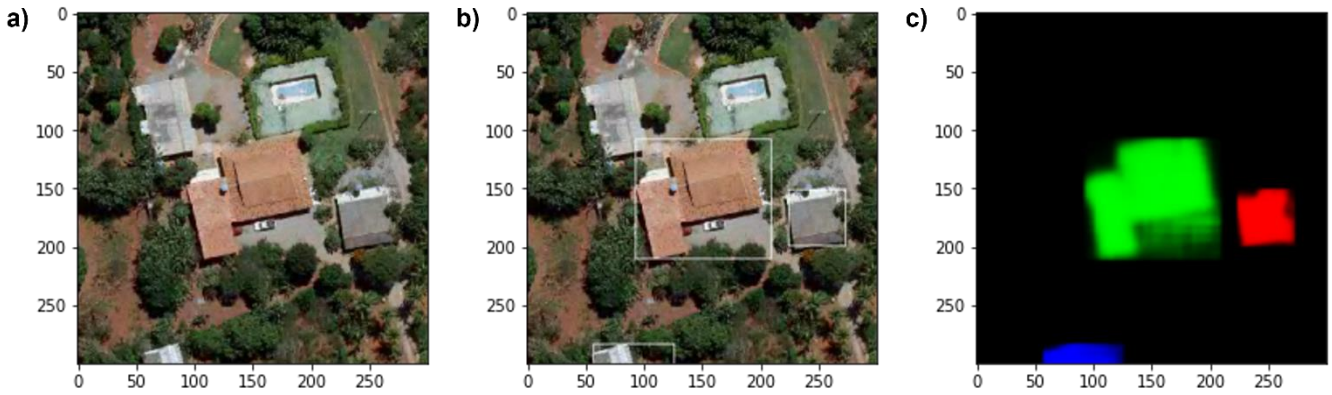

Figura 5. Exemplo de resultado do modelo para detecções com probabilidade mínima de 0,9, sobre uma imagem do *dataset* teste (a), resultando na detecção (b) e segmentação (c) expostas.

Média e Revocação Média, referentes ao enquadramento, e 0,522 e 0,599 referentes à segmentação. Tem-se através da Figura 5 um exemplo de detecções com 0,9 ou mais de probabilidade de conter um objeto de interesse, utilizando uma imagem do *dataset* de teste. A intensidade do pixel aponta para a probabilidade daquele pixel de conter uma detecção. O modelo em questão passou por poucas iterações em seu processo de treinamento devido à limitação já mencionada, portanto, para melhor desempenho, seria razoável mais iterações para melhor ajuste dos parâmetros e consequente maior precisão.

Para se contornar o problema enfrentado pelo modelo que sofreu sobreajuste, sugere-se experimentar o uso da banda espectral do infravermelho próximo como fonte adicional de informação, por apresentar melhor separação entre vegetação e outras classes temáticas como área urbana, por exemplo. Além disto, técnicas de super-resolução, capazes de executar redimensionamento da imagem sem perda de qualidade, representam alternativas otimistas em situações como a abordada neste artigo [3].

4. CONCLUSÕES

Conforme os espaços urbanos tornam-se cada vez mais dinâmicos e heterogêneos, técnicas e metodologias de análise da superfície terrestre devem entrar em conformidade com a nova realidade, visando abarcar também os constantes avanços tecnológicos. Desta forma este artigo buscou a avaliação de uma metodologia emergente de análise para a realidade brasileira e concluiu que esta não se adequa bem, nos termos especificados neste texto, para a tarefa de detecção de edificações utilizando imagens do CBERS-4A.

Após adaptação do modelo e experimentação com diferente hiperparâmetros, parâmetros de treinamento e outras estratégias, o processo de treinamento continuou mostrando aspectos de sobreajuste. A gama de estratégias possíveis não foi exaurida, como discutido na seção anterior.

Não obstante, o treinamento do modelo avaliado em imagens de maior resolução espacial apresentou convergência durante todo o processo, resultando nas métricas de precisão discutidas. Por meio dos resultados apresentados na Figura 5 pode-se observar resultados promissores em relação tanto à detecção como à segmentação, detectando com qualidade satisfatória os limites das edificações.

5. REFERÊNCIAS

[1] D. Ventura, A. Bonifazi, M. F. Gravina, A. Belluscio, and G. Ardizzone, “Mapping and classification of ecologically sensitive marine habitats using unmanned aerial vehicle (UAV) imagery and Object-Based Image Analysis (OBIA),” MDPI AG, 2018.

[2] S. Li, Z. Zhu, H. Wang, and F. Xu, “3D Virtual Urban Scene Reconstruction from a Single Optical Remote Sensing Image,” *IEEE Access*, vol. 7, pp. 68305–68315, 2019.

[3] L. Courtrai, M. T. Pham, and S. Lefèvre, “Small object detection in remote sensing images based on super-resolution with auxiliary generative adversarial networks,” *Remote Sens (Basel)*, vol. 12, no. 19, pp. 1–19, 2020.

[4] G. Cheng and J. Han, “A Survey on Object Detection in Optical Remote Sensing Images,” *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 117, pp. 11–28, 2016.

[5] W. Burger and M. J. Burge, *Digital Image Processing*, 79th ed., vol. 79. 2021.

[6] S. P. Mohanty *et al.*, “Deep Learning for Understanding Satellite Imagery: An Experimental Survey,” *Front Artif Intell*, vol. 3, 2020.

[7] K. He, G. Gkioxari, P. Dollár, and R. Girshick, “Mask R-CNN,” *IEEE International Conference on Computer Vision (ICCV)*, pp. 2980–2988, 2017.

[8] K. He, X. Zhang, S. Ren, and J. Sun, “Deep Residual Learning for Image Recognition,” *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016.

[9] S. Ren, K. He, R. Girshick, and J. Sun, “Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks,” *IEEE Trans Pattern Anal Mach Intell*, vol. 39, pp. 1137–1149, 2017.