Detecção e segmentação de instâncias em tempo real utilizando detectron2: uma análise computacional baseada em redes neurais convolucionais

Vitor Amadeu Souza[1]; 0009-0002-1857-6799

1 – UniFOA, Centro Universitário de Volta Redonda, Volta Redonda, RJ.
vitor.amadeu@foa.org.br

Resumo: Este trabalho apresenta uma análise da aplicação do framework Detectron2 para detecção e segmentação de objetos em imagens urbanas complexas. A pesquisa implementou o modelo Mask R-CNN com backbone ResNet-50 e Feature Pyramid Network (FPN) para identificação automática de múltiplas categorias de objetos, incluindo veículos, pessoas, semáforos e aeronaves. Os resultados demonstraram precisão na detecção com scores de confiança superiores a 90% para a maioria dos objetos identificados. O estudo evidenciou a eficácia do framework em cenários urbanos complexos, destacando sua capacidade de processamento em tempo real e robustez na identificação de objetos sobrepostos. A metodologia empregada utilizou configurações pré-treinadas no dataset COCO, com threshold de confiança ajustado para 0.5, permitindo detecções precisas mesmo em condições de alta densidade de objetos. Os achados contribuem para o avanço das tecnologias de visão computacional aplicadas a sistemas de monitoramento urbano inteligente e veículos autônomos.

Palavras-chave: Detectron2. Mask R-CNN. Detecção de Objetos. Segmentação de Instâncias. Visão Computacional

INTRODUÇÃO

A detecção automática de objetos em imagens tem se tornado uma área de crescente importância na ciência da computação, especialmente com o advento das redes neurais convolucionais profundas. Segundo LeCun, Bengio e Hinton (2015), o aprendizado profundo revolucionou o campo da visão computacional, permitindo avanços significativos em tarefas como classificação, detecção e segmentação de imagens. O desenvolvimento de frameworks especializados, como o Detectron2, representa um marco na democratização de tecnologias avançadas de detecção de objetos para a comunidade científica e industrial.

O Detectron2, desenvolvido pela Meta AI Research (Wu *et al*., 2019), constitui uma reimplementação completa do framework original Detectron, oferecendo melhorias substanciais em termos de flexibilidade, modularidade e performance. Esta plataforma integra algoritmos estado-da-arte como Mask R-CNN (Kantor *et al*., 2020), Faster R-CNN (Ren *et al*., 2015) e RetinaNet (Lin *et al*., 2017), proporcionando uma base sólida para pesquisas em detecção e segmentação de instâncias. A arquitetura modular do framework permite a fácil experimentação com diferentes backbones, cabeças de detecção e estratégias de treinamento, facilitando tanto a pesquisa acadêmica quanto aplicações industriais.

A importância da detecção automática de objetos transcende os aspectos puramente técnicos, apresentando aplicações práticas em diversos domínios. Redmon *et al.* (2016) destacam as aplicações em veículos autônomos, onde a detecção precisa de pedestres, veículos e obstáculos é fundamental para a segurança. Similarmente, Girshick (2015) enfatiza a relevância em sistemas de vigilância inteligente, análise de tráfego urbano e robótica móvel. O processamento em tempo real de imagens urbanas complexas, contendo múltiplos objetos sobrepostos e em diferentes escalas, representa um desafio computacional significativo que tem motivado extensas pesquisas na área.

A segmentação de instâncias, em particular, oferece vantagens substanciais em relação à detecção tradicional baseada em caixas delimitadoras. Long, Shelhamer e Darrell (2015) demonstraram que a segmentação pixel-wise permite uma compreensão mais refinada da cena, essencial para aplicações que requerem precisão espacial elevada. Esta capacidade

é especialmente relevante em cenários urbanos, onde objetos de diferentes categorias frequentemente se sobrepõem ou apresentam oclusões parciais, demandando algoritmos capazes de distinguir fronteiras precisas entre instâncias.

O presente estudo visa avaliar a eficácia do framework Detectron2 na detecção e segmentação de objetos em imagens urbanas complexas, utilizando uma abordagem quantitativa baseada em métricas de confiança e análise qualitativa dos resultados visuais. A pesquisa contribui para o corpo de conhecimento em visão computacional aplicada, fornecendo insights sobre a performance de modelos pré-treinados em cenários reais e demonstrando a viabilidade de implementações práticas em sistemas de monitoramento urbano.

MÉTODOS

A metodologia empregada neste estudo baseou-se na implementação do framework Detectron2 utilizando o modelo Mask R-CNN com backbone ResNet-50 e Feature Pyramid Network. A escolha desta arquitetura fundamentou-se nos trabalhos de He *et al.* (2016), que demonstraram a superioridade das redes residuais profundas na extração de características hierárquicas, e Lin *et al*. (2017), que estabeleceram a eficácia das Feature Pyramid Networks na detecção de objetos em múltiplas escalas.

O threshold de confiança foi estabelecido em 0.5, valor que segundo Huang *et al*. (2017) oferece um equilíbrio adequado entre precisão e recall na detecção de objetos. Esta configuração permite a filtragem de detecções com baixa confiança, reduzindo falsos positivos enquanto mantém a sensibilidade para detecções verdadeiras. O modelo pré-treinado foi carregado automaticamente através do model zoo do Detectron2, eliminando a necessidade de treinamento adicional e permitindo avaliação imediata da performance.

A imagem de teste selecionada representa uma cena urbana complexa contendo múltiplas categorias de objetos, incluindo veículos terrestres, pessoas, infraestrutura urbana e uma aeronave. Esta escolha metodológica baseou-se na necessidade de avaliar a robustez do modelo em cenários realísticos com alta densidade de objetos e potenciais oclusões. A imagem foi pré-processada utilizando técnicas padrão de normalização e redimensionamento, conforme especificado nas configurações do modelo.

O processamento da imagem seguiu o pipeline padrão do Detectron2, iniciando com a extração de características através do backbone ResNet-50, seguido pela geração de propostas de regiões através da Region Proposal Network (RPN) e finalizado com a classificação e refinamento das detecções através das cabeças de detecção especializadas. Girshick *et al.* (2014) descreveram os fundamentos teóricos deste processo, enquanto Ren *et al*. (2015) estabeleceram as bases da arquitetura de duas etapas empregada.

A visualização dos resultados foi implementada utilizando a classe Visualizer nativa do Detectron2, que sobrepõe as máscaras de segmentação e caixas delimitadoras na imagem original. Esta abordagem permite análise qualitativa imediata dos resultados, facilitando a identificação de sucessos e limitações do modelo. As métricas de confiança foram extraídas automaticamente para cada detecção, permitindo análise quantitativa da certeza do modelo em cada predição.

O código-fonte está disponível para download através do link: https://github.com/vitor-souza-ime/detectron2.

RESULTADOS E DISCUSSÃO

Os resultados obtidos através da aplicação do framework Detectron2 na imagem urbana demonstraram performance na detecção e segmentação de múltiplas categorias de objetos. A análise quantitativa revelou scores de confiança consistentemente elevados, com a maioria das detecções superando o threshold de 90% de certeza. Especificamente, os semáforos foram detectados com confiança de 99% e 100%, evidenciando a robustez do modelo na identificação de objetos de infraestrutura urbana, resultado consistente com os achados de Bochkovskiy, Wang e Liao (2020) em suas avaliações de detectores modernos. A Figura 1 demonstra o resultado do processamento do detectron2.

Figura 1 - Detecção de objetos pelo detectron2

Fonte: Wikimedia Foundation

A detecção de pessoas na cena apresentou variabilidade interessante nos scores de confiança, com valores oscilando entre 79% e 100%. Esta variação pode ser atribuída a fatores como oclusão parcial, escala dos objetos e qualidade da imagem na região específica, fenômenos documentados por Dollar *et al*. (2012) em seus estudos sobre detecção de pedestres. Notavelmente, o modelo demonstrou capacidade de distinguir entre indivíduos parcialmente oclusos, mantendo alta precisão mesmo em condições desafiadoras de sobreposição e densidade populacional elevada.

A detecção automobilística revelou scores elevados, com veículos identificados com confiança superior a 98%. Esta performance alinha-se com os resultados reportados por Redmon e Farhadi (2018) em suas avaliações de detectores em tempo real. A precisão na delimitação das fronteiras veiculares e a capacidade de distinção entre diferentes tipos de

veículos (carros, motocicletas) demonstra a sofisticação das características aprendidas pelo backbone ResNet-50, confirmando as observações de He *et al*. (2016) sobre a eficácia das arquiteturas residuais profundas.

Particularmente notável foi a detecção da aeronave com 62% de confiança, um resultado significativo considerando que objetos aéreos em imagens urbanas representam um desafio computacional devido à sua raridade relativa nos dados de treinamento. Lin *et al.* (2014) observaram em sua análise do dataset COCO que a categoria "airplane" apresenta menor representação comparada a objetos urbanos comuns, o que poderia explicar a confiança relativamente menor. Contudo, a capacidade do modelo de identificar corretamente este objeto atípico demonstra a generalização efetiva do aprendizado adquirido durante o treinamento.

A qualidade das máscaras de segmentação geradas pelo componente Mask R-CNN mostrou-se precisa, com fronteiras bem definidas e sobreposições mínimas entre diferentes instâncias. Esta precisão pixel-wise é fundamental para aplicações que requerem localização espacial exata, como navegação robótica ou análise de tráfego avançada. Kirillov *et al*. (2019) enfatizaram a importância desta precisão em seus trabalhos sobre segmentação panorâmica, destacando como máscaras precisas contribuem para compreensão cênica mais refinada.

CONCLUSÕES

A capacidade do modelo de detectar simultaneamente múltiplas categorias de objetos, desde infraestrutura urbana comum até objetos atípicos como aeronaves, demonstra a robustez e generalização do aprendizado adquirido através do treinamento no dataset COCO. Esta versatilidade é fundamental para aplicações reais, onde a diversidade de objetos e cenários requer adaptabilidade computacional elevada.

Pesquisas futuras podem focar na avaliação de performance em condições ambientais adversas, incluindo variações climáticas, iluminação noturna e cenários de alta velocidade. Adicionalmente, estudos comparativos com arquiteturas alternativas e análises de sensibilidade a diferentes thresholds de confiança poderão fornecer insights adicionais para otimização de sistemas práticos.

REFERÊNCIAS

BOCHKOVSKIY, A.; WANG, C. Y.; LIAO, H. Y. M. YOLOv4: Optimal Speed and Accuracy of Object Detection. arXiv preprint arXiv:2004.10934, 2020. Disponível em: https://arxiv.org/abs/2004.10934. Acesso em: 30 ago. 2025.

DOLLAR, P. et al. Pedestrian detection: An evaluation of the state of the art. IEEE Transactions on Pattern Analysis and Machine Intelligence, v. 34, n. 4, p. 743-761, 2012. DOI: 10.1109/TPAMI.2011.155. Disponível em: https://ieeexplore.ieee.org/document/5975165. Acesso em: 30 ago. 2025.

GIRSHICK, R. Fast R-CNN. Proceedings of the IEEE International Conference on Computer Vision, p. 1440-1448, 2015. DOI: 10.1109/ICCV.2015.169. Disponível em: https://ieeexplore.ieee.org/document/7410526. Acesso em: 30 ago. 2025.

GIRSHICK, R. et al. Rich feature hierarchies for accurate object detection and semantic segmentation. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, p. 580-587, 2014. DOI: 10.1109/CVPR.2014.81. Disponível em: https://ieeexplore.ieee.org/document/6909475. Acesso em: 30 ago. 2025.

HE, K. et al. Deep residual learning for image recognition. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, p. 770-778, 2016. DOI: 10.1109/CVPR.2016.90. Disponível em: https://ieeexplore.ieee.org/document/7780459. Acesso em: 30 ago. 2025.

KANTOR, Charles et al. Over-CAM: Gradient-based localization and spatial attention for confidence measure in fine-grained recognition using deep neural networks. 2020.

HUANG, Jonathan et al. Speed/accuracy trade-offs for modern convolutional object detectors. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2017. p. 7310-7311.

KIRILLOV, Alexander et al. Panoptic segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019. p. 9404-9413.

LECUN, Y.; BENGIO, Y.; HINTON, G. Deep learning. Nature, v. 521, n. 7553, p. 436-444, 2015. DOI: 10.1038/nature14539. Disponível em: https://www.nature.com/articles/nature14539. Acesso em: 30 ago. 2025.

LIN, T. Y. et al. Feature pyramid networks for object detection. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, p. 2117-2125, 2017. DOI: 10.1109/CVPR.2017.106. Disponível em: https://ieeexplore.ieee.org/document/8099589. Acesso em: 30 ago. 2025.

LIN, T. Y. et al. Focal loss for dense object detection. Proceedings of the IEEE International Conference on Computer Vision, p. 2980-2988, 2017. DOI: 10.1109/ICCV.2017.324. Disponível em: https://ieeexplore.ieee.org/document/8237586. Acesso em: 30 ago. 2025.

LIN, T. Y. et al. Microsoft COCO: Common objects in context. European Conference on Computer Vision, p. 740-755, 2014. DOI: 10.1007/978-3-319-10602-1_48. Disponível em: https://link.springer.com/chapter/10.1007/978-3-319-10602-1_48. Acesso em: 30 ago. 2025.

LONG, J.; SHELHAMER, E.; DARRELL, T. Fully convolutional networks for semantic segmentation. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, p. 3431-3440, 2015. DOI: 10.1109/CVPR.2015.7298965. Disponível em: https://ieeexplore.ieee.org/document/7298965. Acesso em: 30 ago. 2025.

REDMON, J. et al. You only look once: Unified, real-time object detection. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, p. 779-788, 2016. DOI: 10.1109/CVPR.2016.91. Disponível em: https://ieeexplore.ieee.org/document/7780460. Acesso em: 30 ago. 2025.

REDMON, J.; FARHADI, A. YOLOv3: An incremental improvement. arXiv preprint arXiv:1804.02767, 2018. Disponível em: https://arxiv.org/abs/1804.02767. Acesso em: 30 ago. 2025.

REN, S. et al. Faster R-CNN: Towards real-time object detection with region proposal networks. Advances in Neural Information Processing Systems, v. 28, p. 91-99, 2015. Disponível em: https://papers.nips.cc/paper/2015/hash/14bfa6bb14875e45bba028a21ed38046-Abstract.html. Acesso em: 30 ago. 2025.

WU, Y. et al. Detectron2. GitHub Repository, 2019. Disponível em: https://github.com/facebookresearch/detectron2. Acesso em: 30 ago. 2025.

WIKIMEDIA FOUNDATION. Wikimedia Commons inclui mais de 100 milhões de arquivos de mídia de uso livre – fotos, áudios e vídeos. Disponível em: Wikimedia Commons. Acesso em: 31 ago. 2025.