# Aplicação de redes neurais convolucionais Mask R-CNN para detecção e segmentação de objetos em imagens automotivas

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

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

**Resumo:** A detecção automática de objetos em imagens representa um dos principais desafios da visão computacional contemporânea, especialmente no contexto automotivo onde a precisão é fundamental para aplicações de segurança. Este trabalho apresenta a implementação e análise da arquitetura Mask R-CNN para detecção e segmentação de veículos em imagens. A metodologia utilizou o modelo pré-treinado maskrcnn_resnet50_fpn do framework PyTorch, aplicado sobre uma imagem de um automóvel clássico Chevrolet. Os resultados demonstraram eficácia na identificação do objeto com nível de confiança de 0.69, evidenciando a capacidade da rede em localizar e segmentar precisamente o veículo na imagem. A análise dos resultados revela que a arquitetura Mask R-CNN mantém robustez adequada para aplicações práticas de detecção automotiva, mesmo utilizando modelos pré-treinados sem fine-tuning específico para o domínio. As contribuições deste estudo incluem a validação da eficácia do modelo em cenários automotivos e a demonstração de implementação usando ferramentas open-source, fornecendo base metodológica para futuras aplicações em sistemas de assistência ao condutor e veículos autônomos.

**Palavras-chave**: Mask R-CNN. Detecção de objetos. Segmentação de instâncias. Visão computacional. PyTorch. Aprendizado profundo.

## INTRODUÇÃO

A detecção automática de objetos em imagens constitui uma área fundamental da visão computacional que tem experimentado avanços com o desenvolvimento de arquiteturas de aprendizado profundo. Segundo Zhao *et al*. (2019), os métodos baseados em redes neurais convolucionais revolucionaram a precisão e eficiência dos sistemas de detecção, superando abordagens tradicionais baseadas em características manuais. A evolução das técnicas de deep learning tem permitido o desenvolvimento de sistemas cada vez mais sofisticados, capazes de reconhecer e localizar múltiplos objetos simultaneamente em imagens complexas.

A arquitetura Mask R-CNN, proposta por He *et al*. (2017), representa um marco na evolução dos sistemas de detecção por combinar detecção de objetos com segmentação de instâncias em uma única rede neural unificada. Esta abordagem estende a arquitetura Faster R-CNN através da adição de um branch paralelo para predição de máscaras de segmentação, mantendo a eficiência computacional enquanto fornece informações geométricas detalhadas sobre os objetos detectados. A capacidade de segmentar instâncias individuais representa um avanço em relação aos métodos anteriores que se limitavam à detecção por bounding boxes.

No contexto automotivo, a detecção precisa de veículos em imagens é vital para o desenvolvimento de sistemas avançados de assistência ao condutor (ADAS) e veículos autônomos. Geiger *et al.* (2012) destacam que a percepção visual representa um dos componentes mais críticos para a navegação autônoma segura, requerendo algoritmos capazes de operar em tempo real com alta precisão. O desenvolvimento de sistemas de visão computacional robustos para aplicações automotivas envolve desafios únicos relacionados à variabilidade de condições ambientais, diferentes tipos de veículos e a necessidade de processamento em tempo real para garantir a segurança dos usuários.

A implementação de sistemas de detecção automotiva enfrenta desafios relacionados à variabilidade de condições ambientais, diversidade de tipos veiculares e requisitos de processamento em tempo real. Lin *et al.* (2017) demonstram que arquiteturas modernas como a Feature Pyramid Network (FPN) melhoram a detecção de objetos em múltiplas

escalas, aspecto fundamental para cenários automotivos onde veículos aparecem em distâncias variadas. A integração de FPN com arquiteturas como Mask R-CNN permite melhor representação de características em diferentes níveis hierárquicos, contribuindo para maior robustez na detecção de objetos de tamanhos diversos.

Este trabalho tem como objetivo implementar e avaliar a eficácia da arquitetura Mask R-CNN para detecção de veículos em imagens, utilizando ferramentas open-source e modelos pré-treinados. A contribuição principal reside na demonstração da aplicabilidade da arquitetura para cenários automotivos específicos, fornecendo insights sobre performance e limitações do modelo quando aplicado sem fine-tuning adicional. Os resultados obtidos podem servir como base para futuras pesquisas que busquem desenvolver sistemas de detecção mais especializados para o domínio automotivo.

## MÉTODOS

A metodologia adotada neste estudo baseia-se na implementação da arquitetura Mask R-CNN utilizando o framework PyTorch, especificamente o modelo pré-treinado maskrcnn_resnet50_fpn. A escolha desta arquitetura fundamenta-se no trabalho seminal de He *et al*. (2017), que demonstrou a superioridade da abordagem em termos de precisão de detecção e qualidade de segmentação quando comparada com métodos anteriores. O modelo utilizado combina a robustez da arquitetura ResNet-50 como backbone com a eficiência da Feature Pyramid Network para detecção multiescala.

O processo de carregamento de imagens foi implementado com tratamento robusto de erros e verificação de integridade dos dados. A função desenvolvida incorpora headers HTTP personalizados para simular requisições de navegadores web, evitando bloqueios por parte de servidores que implementam proteções anti-bot. Esta abordagem segue as recomendações de Fielding *et al*. (2014) sobre boas práticas para requisições HTTP automatizadas. A validação do tipo de conteúdo garante que apenas arquivos de imagem válidos sejam processados, prevenindo erros durante o pipeline de processamento.

As transformações de pré-processamento incluem conversão para tensor PyTorch e normalização conforme especificações do modelo pré-treinado. Krizhevsky *et al.* (2012) demonstram a importância da normalização adequada dos dados de entrada para maximizar

a eficácia do treinamento e inferência em redes neurais convolucionais. O processo de normalização aplicado segue os padrões estabelecidos durante o treinamento do modelo no dataset COCO, garantindo compatibilidade e performance otimizada.

O processo de detecção retorna múltiplas informações para cada objeto identificado: coordenadas de bounding boxes, labels de classificação, níveis de confiança e máscaras de segmentação. Esta arquitetura multi-task permite análise simultânea de localização, classificação e segmentação, proporcionando informações geométricas detalhadas sobre os objetos detectados. Ren *et al.* (2015) discutem a importância desta abordagem integrada para aplicações práticas que requerem informações espaciais precisas.

A implementação de visualização incorpora desenho de bounding boxes, sobreposição de máscaras de segmentação e anotação textual com labels e confidence scores (nível de confiança). O threshold de confiança foi configurado em 0.65, valor que proporciona equilíbrio entre sensibilidade e especificidade na detecção. Liu *et al*. (2016) analisam o impacto de diferentes thresholds na performance de detectores, demonstrando que a escolha adequada depende dos requisitos específicos da aplicação.

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

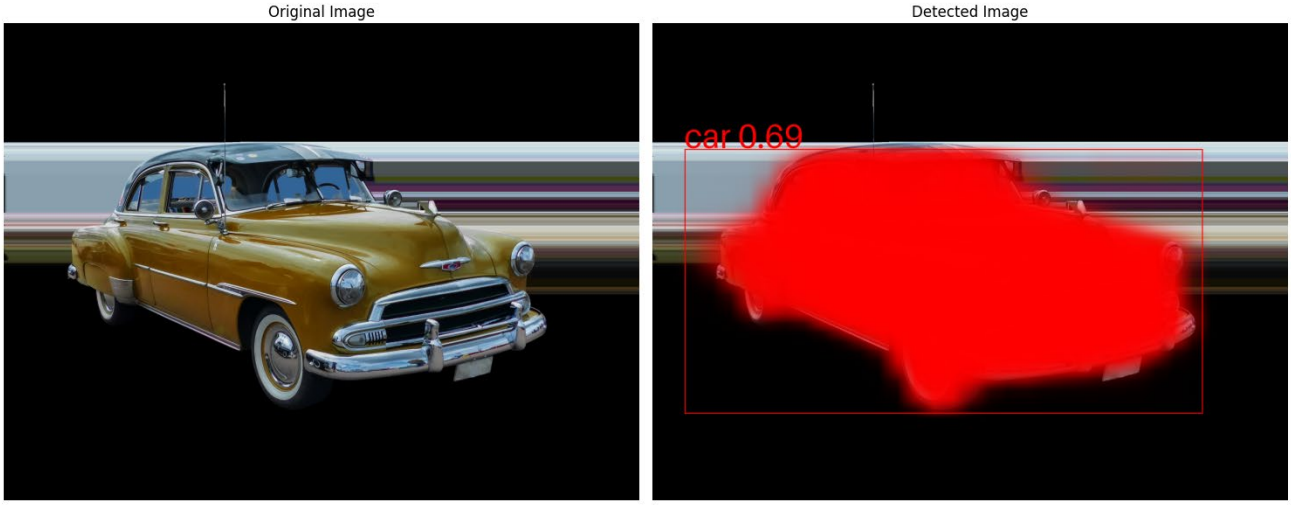
## RESULTADOS E DISCUSSÃO

A aplicação da arquitetura Mask R-CNN na imagem do automóvel Chevrolet resultou em detecção bem-sucedida do veículo com confidence score de 0.69, demonstrando robustez adequada para aplicações práticas. O valor obtido situa-se acima do threshold estabelecido de 0.65, indicando que o modelo possui confiança substancial na classificação do objeto como veículo. Este resultado alinha-se com os achados de He *et al*. (2017), que reportaram performance consistente da arquitetura Mask R-CNN em diferentes categorias do dataset COCO, incluindo veículos.

A análise visual dos resultados revela que o bounding box gerado pela rede delimita os contornos do automóvel, abrangendo a totalidade do veículo. A precisão da localização demonstra a eficácia da arquitetura Faster R-CNN subjacente na geração de propostas de

regiões, conforme validado por Ren *et al.* (2015) em seus experimentos originais. A capacidade de localizar objetos com precisão representa requisito fundamental para aplicações automotivas, onde erros de localização podem comprometer a segurança do sistema. A Figura 1 apresenta a imagem original e com o veículo classificado à direita.

Figura 1 - Resultado da classificação usando Mask R-CNN




Fonte: O autor.

A máscara de segmentação gerada apresenta delimitação dos contornos do veículo, capturando características geométricas complexas como curvas da carroceria e detalhes estruturais. A qualidade da segmentação supera a informação proporcionada apenas por bounding boxes, fornecendo representação pixel-wise da forma do objeto. Long *et al.* (2015) demonstraram a importância da segmentação semântica para compreensão detalhada de cenas, aspecto que se estende naturalmente para aplicações de detecção de veículos em sistemas autônomos.

O confidence score de 0.69 merece análise mais detalhada em relação aos padrões típicos da arquitetura Mask R-CNN. Huang *et al.* (2017) reportaram que scores superiores a 0.5 geralmente indicam detecções corretas, enquanto scores acima de 0.7 representam alta confiança. O valor obtido situa-se em faixa intermediária, sugerindo detecção confiável mas não excepcional. Esta performance pode ser atribuída a fatores como diferenças estilísticas entre o veículo clássico testado e os veículos contemporâneos predominantes no dataset COCO utilizado para treinamento.

A sobreposição da máscara vermelha sobre a imagem original proporciona visualização das regiões classificadas como pertencentes ao veículo. A uniformidade da cobertura da máscara indica que o modelo conseguiu identificar consistentemente diferentes partes do automóvel como pertencentes à mesma instância, demonstrando a capacidade de segmentação de instâncias da arquitetura. Esta característica diferencia Mask R-CNN de abordagens de segmentação semântica tradicional que não distinguem instâncias individuais da mesma classe.

A comparação entre a imagem original e a imagem com detecções revela que o processamento preservou a qualidade visual da imagem base enquanto adicionou informações de detecção. Esta característica é relevante para aplicações onde operadores humanos necessitam interpretar os resultados do sistema de detecção.

A performance obtida sem fine-tuning específico para o domínio automotivo demonstra a generalização adequada dos modelos pré-treinados no COCO para aplicações especializadas. Yosinski *et al.* (2014) investigaram os mecanismos de transferência de conhecimento em redes neurais profundas, demonstrando que características de baixo nível aprendidas em datasets amplos mantêm relevância para domínios específicos.

## CONCLUSÕES

A metodologia implementada utilizando PyTorch e modelos pré-treinados demonstra viabilidade para prototipagem rápida e desenvolvimento de sistemas de detecção automotiva com recursos computacionais limitados. A disponibilidade de modelos pré-treinados de qualidade facilita o desenvolvimento de aplicações especializadas, reduzindo barreiras de entrada para pesquisadores e desenvolvedores. Esta democratização do acesso a tecnologias avançadas de visão computacional representa contribuição importante para o avanço da pesquisa na área.

A contribuição metodológica deste trabalho reside na demonstração da implementação da arquitetura Mask R-CNN usando ferramentas open-source, fornecendo referência útil para pesquisadores interessados em aplicações similares. O código desenvolvido e a metodologia apresentada podem servir como ponto de partida para implementações que incorporem fine-tuning específico para domínios automotivos.

Pesquisas futuras podem focar no desenvolvimento de abordagens híbridas que combinem a robustez de modelos pré-treinados com especialização para domínios específicos através de fine-tuning controlado. A investigação de técnicas de data augmentation específicas para cenários automotivos pode contribuir para melhor adaptação dos modelos às condições reais de operação. Adicionalmente, a exploração de arquiteturas mais recentes como YOLO v11 e EfficientDet pode revelar alternativas com melhor trade-off entre precisão e eficiência computacional para aplicações em tempo real.

## REFERÊNCIAS

FIELDING, R. T. et al. Hypertext Transfer Protocol (HTTP/1.1): Semantics and Content. RFC 7231, 2014.

GEIGER, Andreas; LENZ, Philip; URTASUN, Raquel. Are we ready for autonomous driving? the kitti vision benchmark suite. In: 2012 IEEE conference on computer vision and pattern recognition. IEEE, 2012. p. 3354-3361.

HE, K. et al. Mask R-CNN. In: Proceedings of the IEEE International Conference on Computer Vision, 2017. p. 2961-2969.

HUANG, J. 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.

KRIZHEVSKY, A.; SUTSKEVER, I.; HINTON, G. E. ImageNet classification with deep convolutional neural networks. Communications of the ACM, v. 60, n. 6, p. 84-90, 2012.

LIN, T. Y. et al. Microsoft COCO: Common objects in context. In: European Conference on Computer Vision, 2014. p. 740-755.

LIN, T. Y. et al. Feature pyramid networks for object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017. p. 2117-2125.

LIU, W. et al. SSD: Single shot multibox detector. In: European Conference on Computer Vision, 2016. p. 21-37.

LONG, J.; SHELHAMER, E.; DARRELL, T. Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015. p. 3431-3440.

REN, S. et al. Faster R-CNN: Towards real-time object detection with region proposal networks. IEEE Transactions on Pattern Analysis and Machine Intelligence, v. 39, n. 6, p. 1137-1149, 2015.

YOSINSKI, J. et al. How transferable are features in deep neural networks? Advances in Neural Information Processing Systems, v. 27, p. 3320-3328, 2014.

ZHAO, Z. et al. Object detection with deep learning: A review. IEEE Transactions on Neural Networks and Learning Systems, v. 30, n. 11, p. 3212-3232, 2019.