Avaliação de dois modelos de redes neurais profundas para detecção de objetos

Evaluation of two deep neural networks models for object detection

Pedro Bueno Rios[1], Erikson Freitas de Morais[2]

RESUMO

A visão computacional é uma área da ciência da computação que estuda algoritmos para extrair informações de imagens. A detecção de objetos é uma das tarefas mais importantes da visão computacional, que consiste em encontrar e identificar objetos em uma imagem. Nos últimos anos, a detecção de objetos por meio de redes neurais convolucionais (CNN's) tem se mostrado o estado da arte. As CNN's são capazes de aprender representações visuais complexas, o que lhes permite detectar objetos de forma precisa e eficiente. Existem dois tipos principais de detectores de objetos baseados em CNN's: detectores de um estágio e detectores de dois estágios. Os detectores de um estágio são mais rápidos, mas menos precisos que os detectores de dois estágios. Assim, este trabalho tem como objetivo implementar dois detectores de objetos de um estágio, treinar esses detectores usando um *dataset* de pessoas com e sem máscara e avaliar os detectores usando duas métricas: mAP que informa a precisão das detecções e velocidade de detecção em milissegundos.
PALAVRAS-CHAVE: Redes Neurais Convolucionais; Detecção de Objetos; Métrica.

ABSTRACT

Computer vision is an area of computer science that studies algorithms for extracting information from images. Object detection is one of the most important tasks in computer vision, which consists of finding and identifying objects in an image. In recent years, object detection using convolutional neural networks (CNN's) has become state-of-the-art. CNN's are able to learn complex visual representations, which allows them to detect objects accurately and efficiently. There are two main types of object detectors based on CNN's: one-stage detectors and two-stage detectors. One-stage detectors are faster but less accurate than two-stage detectors. Thus, this work aims to implement two one-stage object detectors, train these detectors using a *dataset* of people with and without masks and evaluate the detectors using two metrics: mAP that report the accuracy of detections and detection speed in milliseconds.
KEYWORDS: Convolution Neural Network; Object Detection; Metrics.

INTRODUÇÃO

Uma única imagem pode conter diversas informações implícitas e explícitas. No entanto extraí-las de forma eficiente é uma tarefa complexa para os computadores atuais, mas em geral, para o cérebro humano não é difícil. Desta forma, a visão computacional é uma subárea da ciência da computação que estuda algoritmos específicos para atacar problemas intrinsecamente visuais, como por exemplo, segmentação, distribuição de cores, estimação de pose e detecção de objetos (PEDRINI;

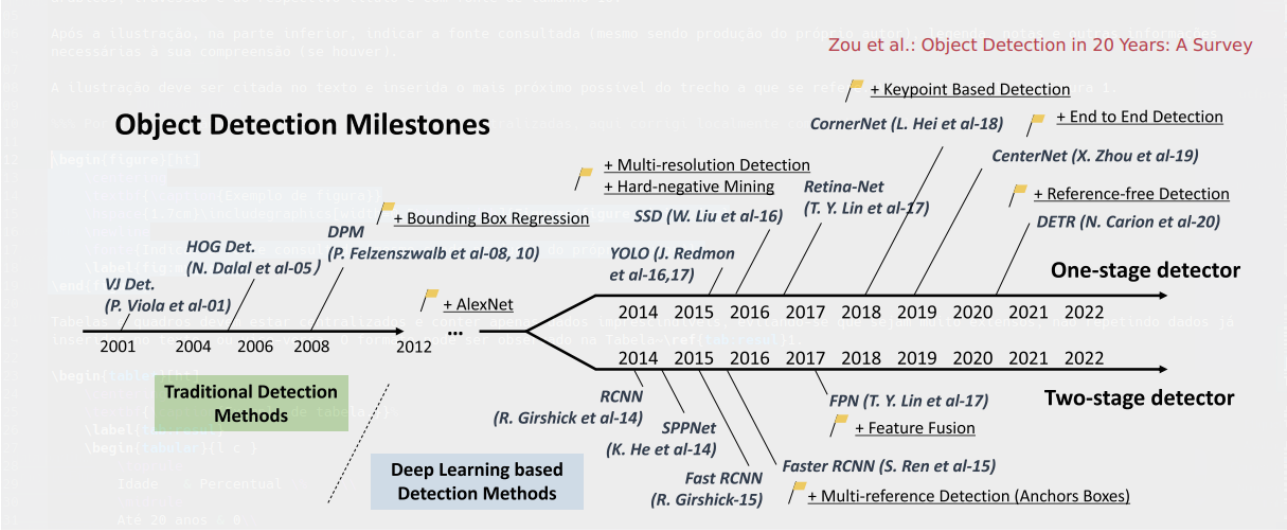
[1] Universidade Tecnológica Federal do Paraná, Ponta Grossa, Paraná, Brasil. E-mail: pedrobuenorios@alunos.utfpr.edu.br. ID Lattes: 9707756035674530.
[2] Docente no Ciência da Computação. Universidade Tecnológica Federal do Paraná, Ponta Grossa, Paraná, Brasil. E-mail: emorais@utfpr.edu.br. ID Lattes: 1716165820460791.

SCHWARTZ, 2007). De acordo com PEDRINI e SCHWARTZ (2007), entende-se que esta área busca facilitar a resolução de problemas complexos simulando a cognição humana e o processo de tomada de decisão, dadas as informações contidas em uma imagem. Assim, é possível observar que a visão computacional possui diversas aplicações no mundo real, como monitoramento por câmeras, biometria, reconhecimento de caracteres, detecção de faces, entre muitas outras (PEDRINI; SCHWARTZ, 2007; ENEMBRECK, 2020; FORSYTH; PONCE, 2002; ULHAQ et al., 2020).

Detectar objetos de interesse consiste em encontrar uma área na imagem de entrada a qual contém pelo menos um objeto de interesse, isto é, de acordo com Zou et al. (2023), determinar técnicas que possam responder uma das peças centrais da visão computacional: Onde estão os objetos e quais são eles?

Para resolver esse desafio, diversas técnicas têm sido desenvolvidas, como a detecção baseada em características visuais (VIOLA; JONES, 2001) e detecção por meio de redes neurais convolucionais (**CNN's**) (KRIZHEVSKY; SUTSKEVER; HINTON, 2012). A Figura 1 apresenta uma linha do tempo dos métodos propostos ao longo dos últimos 21 anos, como Viola e Jones (2001), Krizhevsky, Sutskever e Hinton (2012), Dalal e Triggs (2005) e Redmon et al. (2016). Essas abordagens combinam métodos tradicionais e técnicas de vanguarda para obter resultados cada vez mais precisos.

Figura 1 – Linha do Tempo das técnicas de detecção de objetos.



Fonte: Retirado de Zou et al. (2023).

Na Figura 1 podemos observar alguns detectores tradicionais localizados no tempo, como Viola e Jones (2001) e Dalal e Triggs (2005), como também detectores da vanguarda baseados em redes neurais convolucionais Krizhevsky, Sutskever e Hinton (2012) e logo em seguida a distinção entre detectores de um estágio Redmon et al. (2016), Lin, Goyal et al. (2018), Carion et al. (2020) e detectores de dois estágios como Girshick et al. (2014), Girshick (2015) e Lin, Dollár et al. (2017).

Dessa forma, identificar qual método utilizar no contexto específico do problema é entender o funcionamento dos diversos modelos existentes. Este trabalho toma como objetivo implementar dois modelos do estado da arte *one-stage detectors*, treiná-los usando o *dataset Mask Wearing Dataset* (NELSON, 2022), avaliar as redes treinadas com duas métricas, mAP (*mean Average*

Precision) e velocidade de detecção em milissegundos. Por fim, com base nessas avaliações será possível comparar os modelos afim de entender os resultados alcançados, auxiliando assim, o processo de escolha do detector que melhor se adéqua ao problema.

MATERIAIS E MÉTODOS

Nesta seção será apresentado o *dataset* escolhido para o treinamento, as configurações de treinamento e de testes, como também as ferramentas usadas para o desenvolvimento. Esta seção está dividida em duas subseções a saber: no primeiro momento apresentamos o *dataset*, ou conjunto de dados usado para o treinamento e teste dos modelos escolhidos e na última subseção, é descrita a configuração dos passos de treinamento e teste que foram usados.

DATASET

Datasets em visão computacional são conjuntos de imagens que são utilizados para treinar, validar e testar algoritmos (MURPHY, 2013). A divisão desses *datasets* em três grupos é importante para evitar o *overfitting*, um problema que ocorre quando um modelo se adapta muito bem aos dados de treinamento, mas não é capaz de generalizar para novos dados.

O *dataset Mask Wearing Dataset* (NELSON, 2022) foi desenvolvido com o objetivo de identificar pessoas usando máscaras médicas. Foram coletadas 1934 imagens, em que 1366 foram destinadas para o treinamento, 368 imagens para validação e 200 imagens para teste. O treinamento permite que o algoritmo aprenda a reconhecer características relevantes para identificar o uso de máscaras. A validação testa a capacidade de generalização e ajuda a ajustar hiperparâmetros. Os hiperparâmetros são valores definidos pelos pesquisadores e programadores que influenciam no desempenho do modelo. Já o teste, avalia o desempenho real do algoritmo em situações do mundo real, determinando sua eficácia na detecção de máscaras médicas. As imagens foram originalmente coletadas por Cheng Hsun Teng da Eden Social Welfare Foundation em Taiwan e posteriormente rotuladas por Nelson (2022) em seu trabalho.

YOLOV7 E YOLOV7X

YOLOv7 é um modelo de detecção de objetos de uma etapa desenvolvido por Wang, Bochkovskiy e Liao (2023). É a versão mais recente do modelo YOLO até o momento, que foi lançado pela primeira vez em 2016 (REDMON et al., 2016). Ele toma como foco o aumento na velocidade e precisão da detecção através da agregação de várias camadas de convolução aumentando o número de características extraídas pelas convoluções e combinando-as de forma a melhorar assim o desempenho do modelo. O YOLOv7x é um modelo escalonado do YOLOv7 com mais camadas de convolução e também maiores, tornando-o mais preciso que YOLOv7, mas com velocidade de detecção reduzida.

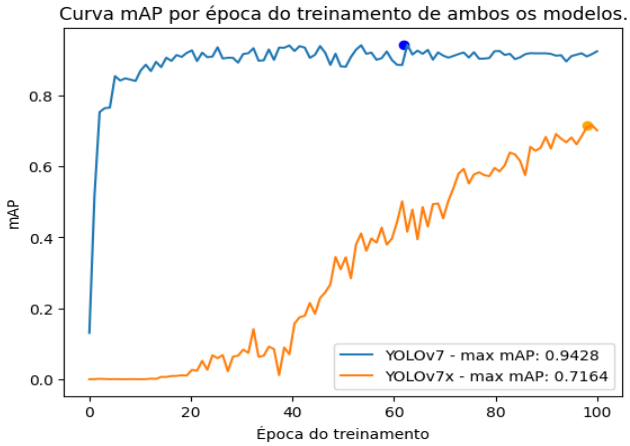
TREINAMENTO E TESTE

Os modelos YOLOv7 e YOLOv7x foram treinados e testados com um *script* Python responsável por executar a rotina de treinamento proposta em (WANG; BOCHKOVSKIY; LIAO, 2023). Este *script* lê um arquivo no formato YAML [1] com as configurações de treinamento e avaliação, ou seja, quantidade de épocas ou iterações do *loop* de treinamento, tamanho do lote que se refere a quantas imagens serão carregadas na GPU, se a rede foi treinada em alguma base de dados anteriormente, estrutura da rede entre outras opções. Com esses dados carregados, o *script* executa a tarefa de treinamento através de uma linha de comando, automatizando o processo de treinamento dos modelos.

Cada modelo foi treinado durante 100 épocas com imagens do tamanho de 480x480x3, 480 pixels de largura e altura mais uma dimensão para as camadas de cores RGB, ou vermelho, verde e azul. Os modelos foram treinados com um total de 1366 imagens.

A avaliação de um modelo em imagens nunca vistas antes é extremamente importante na área de detecção de objetos, visto que é nesse momento que observamos a capacidade de generalização do algoritmo. Dessa forma, na avaliação das redes, o tamanho das 200 imagens para teste foi de 480x480, o mesmo tamanho das imagens utilizada no treinamento. Este teste foi realizado com o objetivo de observar o comportamento das redes e as métricas como velocidade e precisão da detecção, que nos indica a capacidade de generalização.

Figura 2 – Curva mAP por época do treinamento de ambos os modelos.



Fonte: Autoria própria.

A Figura 2 demonstra a variação do mAP ao longo das 100 épocas de ambos os modelos. É possível observar que o YOLOv7, em azul, alcançou um percentual de 94.2%mAP na época 62 enquanto que o YOLOv7x, em laranja, atingiu o máximo de 71.6%mAP na época 98.

EXPERIMENTOS COM YOLOV7 E YOLOV7X

Para podermos avaliar os resultados dos testes propostos, a métrica *mAP* foi aplicada tanto no treinamento quanto na validação das redes, mostrando que na detecção de objetos, é importante

[1] YAML é uma sigla para "YAML Ain't Markup Language"que significa "YAML não é uma linguagem de marcação", frequentemente usado para armazenar configurações .

saber tanto a classificação como também a localização do objeto de interesse. Desta forma, medidas de desempenho como precisão ou *recall* não satisfazem por completo a verificação de qualidade do modelo (EVERINGHAM et al., 2009). As detecções para a avaliação dos modelos foram feitas com um lote de tamanho um, isto é, apenas uma imagem é carregada por vez na GPU.

Tabela 1 – Tabela das métricas de cada modelo apresentado.

Modelo	mAP	Velocidade
YOLOv7	88.8%	41ms
YOLOv7x	71%	65.4ms

Fonte: Autorio própria

A Tabela 1 demonstra as métricas adquiridas após as detecções nas imagens do grupo de teste, como mAP e velocidade de detecção em milissegundos. A partir disso, observa-se que o modelo YOLOv7 atingiu melhor precisão na sua detecção com um mAP de 88.8% enquanto que o YOLOv7x atingiu apenas 71% com tempo de inferência de 24ms maior.

CONCLUSÃO

Por fim é interessante notar que ambos os modelos apresentaram medidas muito boas de localização espacial e classificação dos objetos. No entanto o modelo YOLOv7 atingiu uma porcentagem acima do YOLOv7x como também apresentou menor tempo de detecção. Aumentar o número de épocas de treinamento para o YOLOv7x pode garantir melhor precisão na detecção, mas a velocidade de detecção não irá ser menor que a do YOLOv7 pois as camadas são mais densas e maiores. Dessa forma correlacionar *mAP* e velocidade de detecção é uma tarefa que deve levar em consideração o contexto do problema, como por exemplo, em veículos autônomos em que as detecções precisam ser rápidas. Por outro lado, em câmeras de vigilância as detecções podem ser mais precisas em detrimento do tempo.

Conflito de interesse

Não há conflito de interesse.

REFERÊNCIAS

CARION, Nicolas et al. **End-to-End Object Detection with Transformers**. [S.l.: s.n.], 2020. arXiv: 2005.12872 [cs.CV].

DALAL, N.; TRIGGS, B. Histograms of oriented gradients for human detection. In: 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05). [S.l.: s.n.], 2005. v. 1, 886–893 vol. 1. DOI: 10.1109/CVPR.2005.177.

ENEMBRECK, Fábia Isabela P. Re-identificação de pessoas em imagens digitais utilizando redes neurais siamesas e triple baseadas em uma rede neural convolucional e um autoencoder. Ponta Grossa, PR, Brasil, 2020.

EVERINGHAM, Mark et al. The PASCAL Visual Object Classes (VOC) Challenge. In.

FORSYTH, David A.; PONCE, Jean. **Computer Vision: A Modern Approach**. [S.l.]: Prentice Hall Professional Technical Reference, 2002. ISBN 0130851981.

GIRSHICK, Ross. **Fast R-CNN**. [S.l.: s.n.], 2015. arXiv: 1504.08083 [cs.CV].

GIRSHICK, Ross et al. **Rich feature hierarchies for accurate object detection and semantic segmentation**. [S.l.: s.n.], 2014. arXiv: 1311.2524 [cs.CV].

KRIZHEVSKY, Alex; SUTSKEVER, Ilya; HINTON, Geoffrey E. ImageNet Classification with Deep Convolutional Neural Networks. In: PEREIRA, F. et al. (Ed.). **Advances in Neural Information Processing Systems**. [S.l.]: Curran Associates, Inc., 2012. v. 25. Disponível em: .

LIN, Tsung-Yi; DOLLÁR, Piotr et al. **Feature Pyramid Networks for Object Detection**. [S.l.: s.n.], 2017. arXiv: 1612.03144 [cs.CV].

LIN, Tsung-Yi; GOYAL, Priya et al. **Focal Loss for Dense Object Detection**. [S.l.: s.n.], 2018. arXiv: 1708.02002 [cs.CV].

MURPHY, Kevin P. **Machine learning : a probabilistic perspective**. Cambridge, Mass. [u.a.]: MIT Press, 2013. ISBN 9780262018029 0262018020. Disponível em: .

NELSON, Joseph. **Mask Wearing Dataset**. [S.l.: s.n.], nov. 2022. visitado em 12/08/2023. Disponível em: .

PEDRINI, Hélio; SCHWARTZ, William R. **Análise de imagens digitais**: Princípios, algoritmos e aplicações. São Paulo, SP, Brasil: Cengage Learning Brasil, 2007. 429 p. ISBN 9788522128365.

REDMON, Joseph et al. You Only Look Once: Unified, Real-Time Object Detection. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). [S.l.: s.n.], 2016. P. 779–788. DOI: 10.1109/CVPR.2016.91.

ULHAQ, Anwaar et al. **Computer Vision For COVID-19 Control: A Survey**. [S.l.: s.n.], 2020. arXiv: 2004.09420 [eess.IV].

VIOLA, P.; JONES, M. Rapid object detection using a boosted cascade of simple features. In: PROCEEDINGS of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001. [S.l.: s.n.], 2001. v. 1, p. i–i. DOI: 10.1109/CVPR.2001.990517.

WANG, Chien-Yao; BOCHKOVSKIY, Alexey; LIAO, Hong-Yuan Mark. YOLOv7: Trainable Bag-of-Freebies Sets New State-of-the-Art for Real-Time Object Detectors. In: PROCEEDINGS of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). [S.l.: s.n.], jun. 2023. P. 7464–7475.

ZOU, Zhengxia et al. Object Detection in 20 Years: A Survey. **Proceedings of the IEEE**, v. 111, n. 3, p. 257–276, 2023. DOI: 10.1109/JPROC.2023.3238524.