RECAM
Revista Camalotes

DETECÇÃO DE OBJETOS EM IMAGENS USANDO FASTER R-CNN

OBJECT DETECTION IN IMAGES USING FASTER R-CNN

DETECCIÓN DE OBJETOS EN IMÁGENES USANDO FASTER R-CNN

Edilene A. Veneruchi de Campos[1]

Odirley Franco de Oliveira[2]

RESUMO: Este trabalho de pesquisa explorou a detecção de objetos em imagens por meio da utilização da rede Faster R-CNN. Nesse contexto, foram apresentados os conceitos fundamentais de inteligência artificial e aprendizado de máquina, em especial, o aprendizado profundo, que permite a detecção de padrões complexos em dados brutos. Foi então proposta a implementação de dois modelos de aprendizado baseados na arquitetura da Faster R-CNN, os quais foram treinados e testados com um conjunto de dados criado com imagens de flores, cada uma classificada em duas classes Margarida e Rosa. A escolha da Faster R-CNN foi feita devido à sua eficiência em reduzir o tempo de treinamento e em melhorar a acurácia das respostas obtidas. Com o propósito de verificar a eficiência dos modelos e encontrar respostas satisfatórias, foram estabelecidos dois tipos principais de treinamento: com uma taxa de aprendizagem fixa e com uma taxa de aprendizagem adaptativa. Considerando os resultados, verificou-se que o primeiro modelo apresentou maior valor do aprendizado. Por outro lado, tal sistema também apresentou uma sensibilidade mais limitada e maior incidência de problemas na detecção da borda das caixas e classificação. Em termos de limitações, pode-se citar o pequeno tamanho do *dataset* utilizado e a pouca variação. Como melhorias para o futuro pretende-se usar *dataset* maior e mais variado, empregar estratégias para definição do valor inicial dos hiperparâmetros, como otimização bayseana, além de avaliar outras arquiteturas, como Mask R-CNN. Assim, espera-se compor uma estrutura básica mais eficiente para realizar, além da detecção, a segmentação de instâncias dentro de imagens.

PALAVRAS-CHAVE: detecção de objetos, Faster R-CNN, redes neurais convolucionais, taxa de aprendizado, transferência de aprendizagem.

ABSTRACT: This research work explored object detection in images using the Faster R-CNN network. In this context, the fundamental concepts of artificial intelligence and machine learning were presented, especially deep learning, which allows the detection of complex patterns in raw data. The implementation of two learning models based on the Faster R-CNN architecture was then proposed, which were trained and tested with a dataset created with images of flowers, each classified into two classes: Daisy and Rose. The choice of Faster R-CNN was made due to its efficiency in reducing training time and

[1] Docente e coordenadora do curso de Análise e Desenvolvimento de Sistemas na Faculdade Insted.
[2] Docente na Faculdade Insted.

improving the accuracy of the responses obtained. In order to verify the efficiency of the models and find satisfactory responses, two main types of training were established: with a fixed learning rate and with an adaptive learning rate. Considering the results, it was found that the first model presented a higher learning value. On the other hand, this system also presented a more limited sensitivity and a higher incidence of problems in detecting the edge of the boxes and classification. In terms of limitations, we can mention the small size of the dataset used and the little variation. As improvements for the future, we intend to use a larger and more varied dataset, employ strategies to define the initial value of the hyperparameters, such as Bayesian optimization, in addition to evaluating other architectures, such as Mask R-CNN. Thus, we hope to compose a more efficient basic structure to perform, in addition to detection, the segmentation of instances within images.

KEYWORDS: object detection, Faster R-CNN, convolutional neural networks, learning rate, transfer learning.

RESUMEN: Este trabajo de investigación exploró la detección de objetos en imágenes mediante el uso de la red Faster R-CNN. En este contexto, se presentaron los conceptos fundamentales de la inteligencia artificial y el aprendizaje automático, en particular, el aprendizaje profundo, que permite la detección de patrones complejos en datos sin procesar. Luego se propuso implementar dos modelos de aprendizaje basados en la arquitectura Faster R-CNN, los cuales fueron entrenados y probados con un conjunto de datos creado con imágenes de flores, cada una clasificada en dos clases Daisy y Rose. La elección de Faster R-CNN se realizó debido a su eficiencia para reducir el tiempo de entrenamiento y mejorar la precisión de las respuestas obtenidas. Para comprobar la eficiencia de los modelos y encontrar respuestas satisfactorias se establecieron dos tipos principales de entrenamiento: con una tasa de aprendizaje fija y con una tasa de aprendizaje adaptativo. Considerando los resultados, se encontró que el primer modelo presentó mayor valor de aprendizaje. Por otro lado, este sistema también presentó una sensibilidad más limitada y una mayor incidencia de problemas en la detección y clasificación de los bordes de las cajas. En términos de limitaciones, podemos mencionar el pequeño tamaño del conjunto de datos utilizado y la poca variación. Como mejoras para el futuro, pretendemos utilizar un conjunto de datos más grande y variado, emplear estrategias para definir el valor inicial de los hiperparámetros, como la optimización baysiana, además de evaluar otras arquitecturas, como Mask R-CNN. Así, se espera componer una estructura básica más eficiente para realizar, además de la detección, la segmentación de instancias dentro de imágenes.

PALABRAS CLAVE: detección de objetos, Faster R-CNN, redes neuronales convolucionales, tasa de aprendizaje, transferencia de aprendizaje.

INTRODUÇÃO

Aprendizagem de Maćquina (*machine learning*) tem-se tornado popular em diversas áreas, como diagnósticos médicos; justiça e segurança pública; recrutamento e seleção de recursos humanos; comércio eletrônico;

agricultura de precisão; dentre outros. Em problemas relacionados a processamento de linguagem natural ou a reconhecimento de imagens, tem sido usado um subconjunto de *machine learning*, chamado Aprendizagem Profunda (*deep learning*). Diferentemente da aprendizagem de máquina convencional, aprendizagem profunda consegue detectar padrões complexos e abstratos, aprendendo representações e recursos úteis a partir de dados brutos. O objetivo deste trabalho é investigar, analisar e testar modelos de redes na detecção de objetos em imagem. As seções a seguir apresentam breve fundamentação teórica, os materiais e métodos empregados, análise de resultados e comentários conclusicos, apontando trabalhos futuros.

FUNDAMENTAÇÃO TEÓRICA

Muito da popularidade de aplicações relacionadas a visão computacional está relacionada com o uso de Redes Neurais Profundas (*Deep Neural Networks* - DNN) e com o paradigma de aprendizagem orientada a dados, que permitem aos modelos aprenderem representações úteis automaticamente a partir dos dados. Esta seção aborda, sucintamente, temas importantes para melhor delinear o trabalho desenvolvido.

REDES NEURAIS E REDES NEURAIS PROFUNDAS

Rede Neural Artificial (RN) é um modelo computacional inspirado nas redes neurais humanas, formado por um conjunto de neurônios conectados em um grafo, que pode ser cíclico (recorrente) ou acíclico (*feedfoward*). A organização dos neurônios, normalmente, se dá em camadas e, assim, um neurônio recebe entradas vindas de neurônios da camada anterior via dendritos (processo de sinapse), realiza o processamento necessário (função de ativação) e envia o resultado para neurônios da camada posterior, através do axônio. RN são formadas por camada de entrada, camada de saída e um conjunto de camadas intermediárias, chamadas camadas ocultas [Faceli

2011].

Aprendizagem profunda está relacionada com redes neurais profundas (*Deep Neural Network* - DNN) que recebem este nome por possuírem muitas camadas ocultas. Uma DNN consegue obter vários níveis de abstração, melhorando o processamento de imagem, vídeo e áudio [Goodfellow 2016].

REDES NEURAIS CONVOLUCIONAIS

Dentro das classes de DNN, existem as Redes Neurais Convolucionais (*Convolutional Neural Networks* - CNN), aplicadas principalmente na análise de imagens em problemas de classificação e de reconhecimento de padrão.

A camada de convolução tem a função de extrair características da imagem através da multiplicação da imagem de entrada por uma matriz de pesos. No processo de convolução existem filtros ou *kernels* que percorrem toda a imagem, multiplicando seus valores pelos valores contidos em uma região específica da imagem. Após realizar a convolução de uma região, o filtro é deslocado para a próxima região na horizontal ou na vertical. O tamanho desse deslocamento é chamado de *stride*.

Depois de várias camadas de convolução e de *pooling*, a matriz de características é convertida em um vetor de características. Com camadas totalmente conectadas, essas características são combinadas na criação de um modelo para classificações. Na camada de saída há a aplicação de uma função de perda para cálculo do erro na previsão. Depois disso, os pesos e bias são atualizados por meio de um algoritmo de *backpropagation* para redução de erros e perdas [Ingargiola 2019].

TRANSFERÊNCIA DE APRENDIZAGEM E AUMENTO DE DADOS

Transferência de aprendizagem é a reutilização do conhecimento já adquirido durante o treinamento de um modelo. Este processo é utilizado para reduzir o esforço necessário que se tem ao treinar um modelo do zero. [Silva

2018], cita que existem 3 formas da transferência de conhecimento a ser realizada: extração de características (*feature extraction*), no *fine tuning*; e no treinamento conjunto (*joint training*).

Aumento de dados é uma técnica utilizada para gerar novos exemplares de dados de treinamento a fim de aumentar a generalidade do modelo. Exemplos desta técnica são transformações geométricas, como rotação, translação, escala e corte, dentre outros.

DETECÇÃO DE OBJETOS

A detecção de objetos é uma das áreas de estudo mais ativas em visão computacional. O propósito da classificação de objetos é separar objetos do fundo da imagem (*backround*) e categorize-los de acordo com critérios previamente definidos. Por localização de objetos, entende-se o desenho de uma caixa delimitadora ao redor do objeto. Com o desenvolvimento de redes neurais convolucionais (CNN), realizar detecção de objetos tem sido notadamente melhorada.

Ocorre que, em muitos casos de detecção de objeto, pode haver necessidade de delinear não uma, mas muitas caixas delimitadoras, que representam diferentes objetos de interesse na mesma imagem, não sendo possível antever essa quantidade. Nesses casos, não é possível empregar CNN padrão seguida por uma camada totalmente conectada, pois o comprimento da camada de saída é variável pois o número de ocorrência de objetos de interesse dentro da imagem não é fixo.

Uma estratégia empregada para resolver tal situação, mantendo a viabilidade em termos de recursos computacionais necessários, chamada *Region with Convolutional Neu- ral Network* (R-CNN), foi proposta em [Girshick et al. 2014]. Este método basicamente se divide em duas etapas: execução do algoritmo Selective Search e classificação das regiões geradas na saída da CNN.

A partir da imagem de entrada são extraídas propostas de região com o algoritmo de pesquisa seletiva. Cada uma dessas regiões são redimensionadas em um quadrado e esse quadrado é fornecido como entrada para uma rede neural convolucional, que, por sua vez, produz um vetor de características como saída. A rede neural convolucional atua como um extrator de características e a camada densa de saída contém os recursos extraídos da imagem. Os recursos são, então, passados como entrada o SVM, que classificará a presença do objeto na proposta de região candidata. Esta estratégia é empregada em redes, como R-FCN [Dai et al. 2016], Fast R-CNN [Girshick 2015].

A rede Faster R-CNN[6] [Ren et al. 2016], propõe alteração no processo. De forma similar à Fast R-CNN, uma imagem inteira é fornecida como entrada para uma rede convolucional responsável por gerar um mapa de características convolucionais. Contudo, a Faster R-CNN não usa o algoritmo de pesquisa seletiva (descrito anteriormente) para a identificação das propostas de região, mas sim, faz uso de uma rede separada (*Region Proposal Network* - RPN) para prever as propostas da região. Conceitualmente, Faster R-CNN é composta por três redes neurais: Rede de características (*Feature Network*); Rede de Proposta de Região (*Region Proposal Network* - RPN); e Rede de Detecção (*Detection Network*).

MATERIAIS E MÉTODOS

As possibilidades de experimentos envolvendo tratamento de imagens são muitas e a quantidade de conjuntos de dados públicos disponíveis também é grande.

CONJUNTO DE DADOS

Considerando que o presente trabalho teve como propósito promover a aprendizagem e o entendimento da utilização de algoritmos de aprendizagem de máquina, optou-se pela construção de um *dataset* próprio para

entendimento do processo de rotulação de cada imagem.

O *dataset* construído foi composto por 56 imagens de flores, categorizadas e duas classes: Margaridas e Rosas (28 imagens de cada classe). O *dataset* utilizado foi criado a partir de imagens disponíveis na web. Foram escolhidas imagens de tamanhos aleatórios, pois a rede Faster R-CNN assume a responsabilidade de aplicar padronização nos tamanhos.

Dentro das imagens foram rotuladas as ocorrências de flores do tipo Margarida e Rosa utilizando o LabelImg[3] (figura 1). Para cada imagem foi gerado um arquivo com a extensão xml, identificando os retângulos onde se encontrava alguma Margarida ou Rosa. Tais arquivos, posteriormente foram unificados em um arquivo csv para compor a base de treinamento e a base de teste.

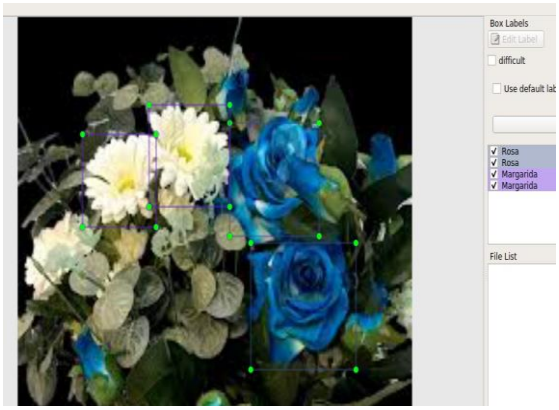

Figura 1. Rotulação dos objetos da imagem realizado com o LabelImg

CONFIGURAÇÕES DO MODELO

Após estudos sobre processamento de objetos em imagens, optou-se pela rede Faster R- CNN, visto que ela emprega redes neurais convolucionais (e consequentemente redes neurais profundas) e algoritmo de classificação como o *Support Vector Machine*. De acordo com [Gir15], essa rede possui dois módulos: 1) uma rede de proposta de região (Region Proposal Network - RPN); e 2) uma rede Fast R-CNN.

[3] https://github.com/tztalin/lablImg

Para realização dos testes, utilizou-se o repositório de modelos de detecção de objetos TensorFlow[4]. O modelo específico empregado foi faster rcnn inception v2 coco[5]. A escolha deste modelo se deu após análise dos resultados publicados do pré-treinamento dos modelos com o *dataset* Coco (*Commom Objetct in Context*)[6], que mostra uma relação diretamente proporcional entre o tempo de treinamento e a acurácia, ou seja, o modelo faster rccn inception v2 coco ficou na região intermediária da tabela, mostrando que durante o pré-treinamento consumiu 58ms e a acurácia obtida foi de 28 mAP.

ETAPA DE TREINAMENTO

O modelo pré-treinado permite a configuração de um conjunto de parâmetros, de acordo com as necessidades. Foi escolhido um modelo de rede pré-treinado como forma de inicialização. Esta estratégia foi estabelecido através da propriedade *fine tune checkpoint*, que recebeu o valor *"faster rcnn inception v2 coco/model.ckpt"*[7]. Com isso, os testes passaram a fazer uso de transferência de conhecimento.

Por se tratar de um conjunto de dados pequeno, também foi utilizada a estratégia de Aumento de Dados. Para isso, a propriedade *data augmentation options* recebeu o valor *"random horizontal flip"*. Como forma de viabilizar futuras comparações, foi esta belecido como *loss* aceitável, valor inferior a 0,05.

Partindo destas configurações básicas, foram executados dois treinamentos. O primeiro utilizou taxa de aprendizagem fixada em 0,0002. O segundo usou taxa de aprendizagem variável, iniciando em 0,002, passando

[4] https://github.com/tensorflow/models/tree/master/research/object-detection
[5] https://github.com/tensorflow/models/blob/master/research/object detection/g3doc/tf1-detection zoo.md
[6] https://cocodataset.org
[7] https://github.com/tensorflow/models/blob/master/research/object detection/g3doc/tf1 detection zoo.md

para 0,0002 a partir do passo 25.000 e para 0,00002 a partir do passo 50.000. O treinamento foi realizado usando ambiente de execução com uma GPU, disponiblizado no Google Colab.

RESULTADOS

No treinamento 1 foram necessários mais de 81.000 passos para que o valor da *loss* ficasse abaixo de 0,05. Nas figuras 2 e 3 é possível acompanhar a evolução *loss* ao longo de todos os passos. Dada a sua arquitetura, a rede Faster R-CNN gera diferentes valores de *loss*.

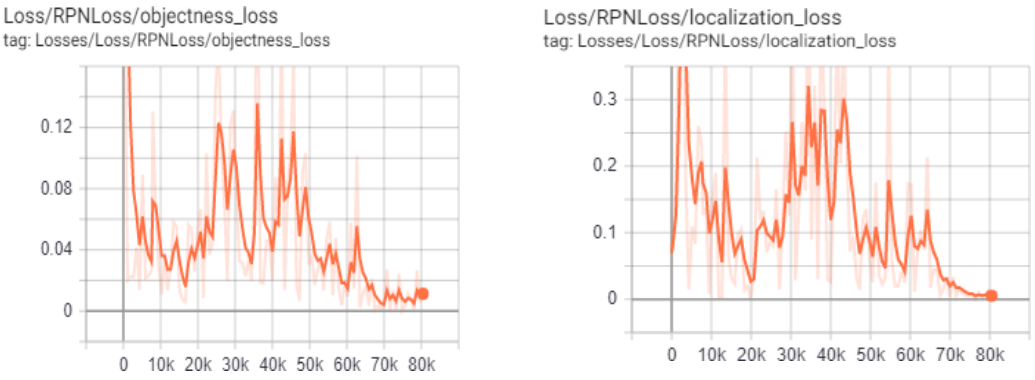

Figura 2. Losses gerados pela RPN

A figura 2 mostra as curvas geradas pelos valores do decorrentes do processo de classificação realizado pela RPN, separando um possível objeto do fundo da imagem. Também mostra as curvas geradas pelos valores da *loss* decorrentes do processo de regressão realizado pela RPN, com o objetivo de propor possíveis caixas circundando o objeto.

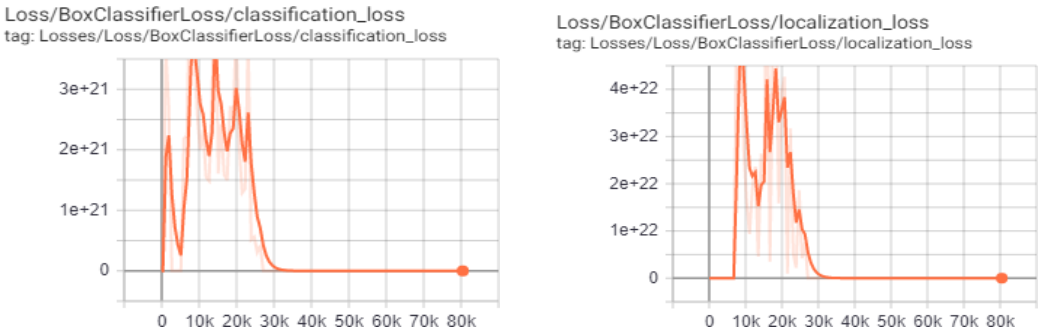

Figura 3. Losses gerados pela Rede de Classificação

A figura 3 mostra a curva gerada pelos valores da *loss* gerados no processo de classificação dos objetos contidos nas caixas delimitadas. Mostra também a curva gerada pelos valores da *loss* decorrentes do processo de regressão para refinar a localização das caixas em torno dos objetos.

No treinamento 2, onde foi utilizada a estratégia de taxa de aprendizagem dinâmica, obteve-se *loss* inferior a 0,05 com 50.000 passos executados, mostrando uma tendência de conversão melhor que a observada no treinamento 1.

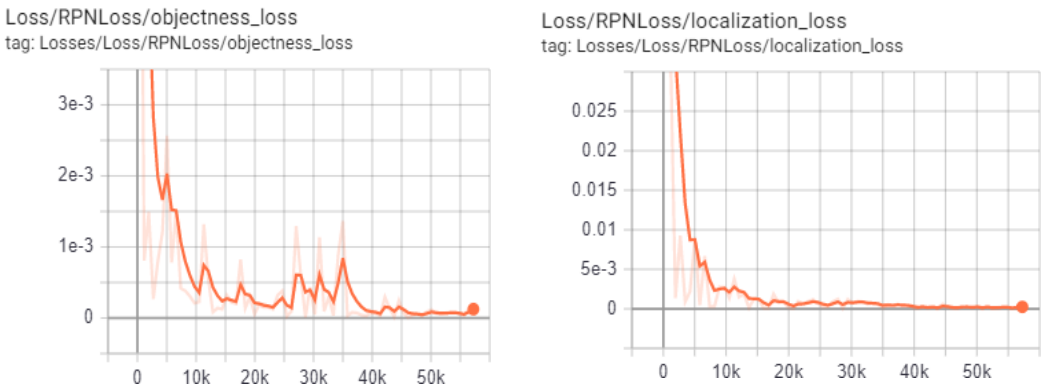

Figura 4. Losses gerados pela RPN

A figura 4 mostra as curvas geradas pelos valores da *loss* decorrentes do processo de classificação realizado pela RPN, separando um possível objeto do fundo da imagem. Também se vê na figura os valores da *loss* decorrentes do processo de regressão realizado pela RPN, com o objetivo de prever possíveis localizações de caixas circundando objetos.

Figura 5. Losses gerados pela Rede de Classificação

A figura 5 mostra as curvas geradas pelos valores do *loss* obtidos no processo de regressã o para refinar a delimitação de caixas em torno dos objetos. Também é possível ver as curvas geradas pelos valores do *loss* decorrentes do processo de classificação dos objetos contidos nas caixas delimitadas.

Conforme mostrado nos gráficos, os valores de *loss* observados no treinamento 2 apontam para convergência, enquanto os do treinamento 1, oscilaram muito, sendo, então, necessária uma quantidade de passos bem maior para fazer os valores para convergência.

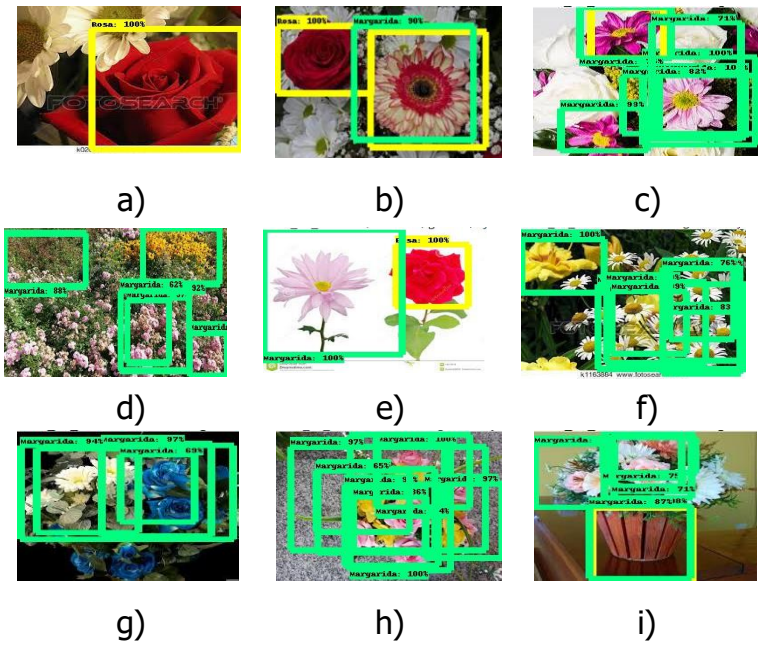

Figura 6. Resultados de testes com o modelo treinado na estratégia 1

Depois do treinamento, foram realizados outros testes com o objetivo de verificar a capacidade de generalização da rede. Os resultados obtidos podem ser melhorados pois, em alguns, ocorreram problemas com a delimitação das caixas em torno dos objetos e em outros ocorreram problemas na classificação.

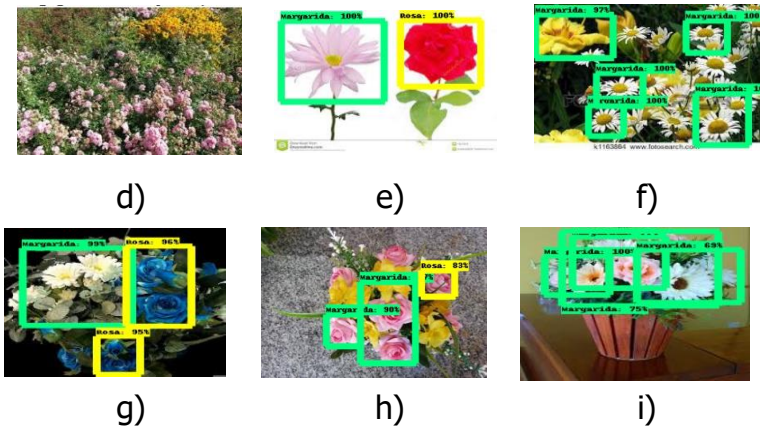

Figura 7. Resultados de testes com o modelo treinado na estratégia 2

Contudo, a comporação entre os resultados gerados pelos modelos, mostra que o uso de taxa de aprendizagem dinâmica é promissor. A figura 6 mostra alguns resultados obtidos nos testes com o modelo do treinamento 1. Alguns exemplos obtidos nos testes com o modelo gerado pelo treinamento 2 são mostrados na figura 7.

CONCLUSÃO

O presente trabalho teve como objetivo estudar estratégias para detecção de objetos em imagens, confrontando-as com a utilização ou não de algoritmos clássicos da área de inteligência artificial.

Para tanto, foi necessária a realização de pesquisa bibliográfica que permitiu conhecer o estado da arte em termos de tratamento de imagens, onde são aplicados algoritmos de busca, redes neurais, redes neurais profundas e redes neurais convolucionais. Estratégias como transferência de aprendizagem e aumento de dados também são amplamente aplicados.

Foram realizados estudos de diferentes arquiteturas para tratamento de imagens, como R-CNN, Fast R-CNN e Faster R-CNN. A Faster R-CNN foi escolhida por ser uma evolução da Fast R-CNN, capaz de diminuir o tempo no processo de treinamento, fornecendo mais acurária nos resultados. Os testes realizados mostraram que existem muitas possibilidades de configuração dos modelos, como forma de melhorar o desempenho.

Nos dois modelos treinados e testados foram empregadas as técnicas de transferência de aprendizagem e de aumento de dados. Para o treinamento 1 utilizouse taxa de aprendizagem fixa e para o treinamento 2 utilizou-se taxa de aprendizagem dinâmica. Nos testes executados após o treinamento, o modelo 2 conseguiu melhor detectar os objetos e traçar as caixas em torno dos objetos com mais acuidade.

Contudo, como todo o trabalho foi realizado com um *dataset* é inevitável que em trabalhos futuros o mesmo procedimento seja repetido com um *dataset* maior. Alé m disso, é importante manipular outros parâmetros no treinamento, além dos aqui descritos, com o objetivo de melhorar os resultados. Por fim, outras arquiteturas podem também ser utilizadas, como a Mask R-CNN, para medir os resultados obtidos com segmentação de instâncias (*instannce segmentation*).

REFERÊNCIAS

Dai, J., Li, Y., He, K., and Sun, J. (2016). **R-fcn: Object detection via region-based fully convolutional networks**.

Faceli, K.; Lorena, A. C. G. J. C. A. (2011). ***Inteligê ncia Artificial: Uma abordagem de aprendizado de má quina"***. Livros Té cnicos e Ciê ntificos Editora Ltda.

Girshick, R. (2015). Fast r-cnn.

Girshick, R., Donahue, J., Darrell, T., and Malik, J. (2014). **Rich feature hierarchies for accurate object detection and semantic segmentation**.

Goodfellow, I.; Bengio, Y. A. (2016). ***DeepLearning*. MITPress**. Ingargiola, A. (2019). **Deep-dive into convolutional networks**.

Ren, S., He, K., Girshick, R., and Sun, J. (2016). **Faster r-cnn: Towards real-time object detection with region proposal networks**.

Silva, L. P. (2018). **Leannet: Uma arquitetura que utiliza o contexto da cena para melhorar o reconhecimento de objetos. Master's thesis**, Pontíficia Universidade Cató lica do Rio Grande do Sul, Porto Alegre.