2D IMAGE OBJECT DETECTION AIDED BY GENERATIVE ADVERSARIAL NETWORKS: A LITERATURE REVIEW

Caio Vinicius Bertolini[a], Roberto Monteiro[a]

[a], Centro Universitário SENAI CIMATEC, Brazil

Abstract: Object Detection (OD) is one of the most important tasks in 2D image processing. Multiple math models have been proposed and frameworks based in Deep Convolutional Networks such as R-CNN, SSD and YOLO are most common. Generative Adversarial Nets (GAN's) represent a prominent field of study in machine learning and it has been applied to many tasks with exciting results. The objective of this work is to assess the potential of GAN's applied to OD tasks and the proposed frameworks as field of study. The methodology used was a systemically review of 14 papers. The conclusion shows that even though OD and GAN's are popular themes, there are not many developments done in the intersection of both subjects. Therefore, OD with GAN applied tasks are an excellent field to explore in future works.

Keywords: Generative Adversarial Nets; Object Detection; Deep Learning

DETECÇÃO DE OBJETOS EM IMAGENS 2D AUXILIADA POR REDES ADVERSÁRIAS GENERATIVAS: UMA REVISÃO DE LITERATURA

Resumo: Detecção de objetos (DO) é uma das tarefas mais importantes dentro do processamento de imagens em 2D. Para isso muitos modelos foram propostos sendo os mais comuns, baseados em redes profundas convolucionais como: R-CNN, SSD e YOLO. As redes generativas adversárias (RGA's) vêm ganhando grande destaque na academia, com aplicações diversas e resultados interessantes. O objetivo deste trabalho é avaliar a aplicação de RGA's em tarefas de DO como potencial área de estudo. A metodologia para isso foi uma revisão de literatura sistêmica de 14 trabalhos. Concluimos que apesar de serem temas bastante populares, não existem grandes desenvolvimentos na intersecção de RGA's e OD. Portanto, um excelente campo a ser explorado em trabalhos futuros e com grades potenciais.

Palavras-chave: Redes Generativas Adversárias; Detecção de Objetos; Aprendizado Profundo de Máquinas

1. INTRODUCTION

Digital Images captured by cameras are very common today as the number of such devices spiked over the last years. The captured images are no longer used only for entertainment or as an art form, they have become a very important source of data that can be analyzed. In such a way that cameras are one of the most common sensors and are present in smartphones, vehicles, smart house devices, cities security systems and many others.

There are many aspects of an image that can be analyzed and one of the most common problem to solve is object detection in 2D images. The object detection task is a combination of other 2 tasks: object classification, which identifies what type of object is being detected, and object localization, which identifies the location the object occupies in the image.

The object detection problem is gaining significant attention from the academic community and getting increased momentum in publications over the last years. Many object detection frameworks have been developed since this problem started being studied such as Viola Jones Detectors, HOG Detector, Deformable Part-based Model (DPM), but Deep Learning techniques, and most specifically the ones based on Convolutional Neural Networks (CNN's), represented a big leap in accuracy and are the most used today. [1]

Today, there are two main categories of object detectors:

1. Two Stage Object Detectors, which separate the detection tasks of classification and localization to be held by different parts of the network

2. Single Stage Object Detectors, which do the classification and localization tasks all at once.

1.1.Two Stage Object Detectors

The Regions with CNN (R-CNN) [2] framework consists of an initial selective search that generates region proposal by close by pixels similarities. Then each proposed region is fed into a CNN for feature extraction. The output is used as input of bound-box regression and classification Support Vector Machine's (SVM's) to define if there is an object in the proposed region and what class it is. The framework presented higher Average Precision (AP) than existing methods for multiple classes [2], but with a high computational cost as it performs redundant computation in the overlapping features of the many proposed regions. [1]

Fast R-CNN [3] and SPPNet [4] apply the CNN only once in the entire image and not many times in the multiple proposed regions as per R-CNN. While Fast R-CNN uses a Region of Interest (RoI) pooling with Fully Connected (FC) layers for classification and bounding box regression [3], SPPNet uses an Spatial Pyramid Pooling (SPP) to define the regions, which also allows for different image input sizes. [4]

Faster R-CNN [5] presents inference results 34 times faster than Fast R-CNN. This happens due to the introduction of Anchor boxes and a specific Fully Convolutional Network (FCN) to generate the Region of Interests (RoI), the Region

Proposed Network (RPN). The RPN uses the Anchor boxes as inputs and Intersection over Union (IoU) metrics to define the RoI bounding boxes, after that it uses the Non-Maximum Suppression (NMS) technique to define a final bounding box. The RPN is trained together with the rest of the model and the region outputs then passes though the RoI pooling and the FC layers for classification and bounding box regression similarly to the Fast R-CNN. [5]

1.2. Single Stage Object Detectors

You Only Look Once (YOLO) [6] method process the entire image by a CNN based model with multiple Anchor Boxes that simultaneously predict the class and location of the bounding boxes. The bounding boxes within a pre-defined threshold are subjected to NMS to define the final bounding box for each object.

Singe Shot Multibox Detector (SSD) [7] also uses multi Anchor boxes. In SSD, after the image passes through the feature extractor, it then goes by Multi-scale feature layers, which decrease in size at each step and allow predictions of detections at multiple scales at once. This ability to detect in multi-resolution and scales is the main contribution of SSD and presented an advantage in average precision when compared to YOLO. The last step of SSD is NMS, that eliminates the overlapping bounding boxes. [1,7]

1.3. Generative Adversarial Networks (GAN's)

GAN [8] is a method of generative network using deep learning. It is divided into two parts below:

The generator, which is an unsupervised model that takes random input of data, usually from Gaussian distribution, and has the objective to deliver a result that can convince a discriminator models that it is real.

And the discriminator, which is a supervised model trained with real data from a dataset and fake data from the generator and has the objective of distinguish one from the other.

Both models are trained together using a loss function defined by the authors as a two-player minmax game below, where G is the generator and D is the discriminator:

$$min_G max_D V(D,G) = E_{x\sim p_{data}(x)}[\log D(x)] + E_{z\sim p_z}[\log\left(1 - D\left(G(z)\right)\right)] \qquad (1)$$

The GAN model is then trained until both, the generator and discriminator models find an equilibrium and the discriminator can no longer identify the difference of data generated by the generator and real data.

GAN's are very exciting field of work and are rapidly evolving and being applied in many different problems and experimentations. [9]

1.4. Objective

This work has the objective of doing a systematic review of published and pre-print works that apply GAN's into image 2D object detection tasks, alone or in combination with existing object detection methods, and compare their methods and frameworks. The results of this review will then be used to assess the viability of GAN's applied to OD tasks and it's potential as a field of study to be applied in future works.

2. METHODOLOGY

To present a comprehensive overview of GAN's application to 2D Image object detection, we searched many academia and Artificial Intelligence community databases such as [10], [11], [12] and [13].

Since it is a rapidly evolving field, many of the papers found are yet to be published and are released as pre-prints. The key words used for the search were: GAN, Object, Detection, Generative, Adversarial.

From the many works found, we first filtered the results by their titles, eliminating all the ones that were not relevant and then by their abstracts. Since the focus of this work is the combination of GAN techniques with object detection tasks, all papers that deviate from that were discarded, including the many papers that use GAN's for data augmentation or 3D object detection. The papers that passed in these criteria were then downloaded from their respective database and read by the authors of this paper.

An analysis was made in the chosen papers with evaluation of the methods and models used and the results obtained.

3. RESULTS AND DISCUSSION

GAN and Object Detection are very popular themes in academy. Initial search in [11] by the key words "generative adversarial" and "object detection" presented 851 results. A Similar process was done in [13] with 98 initial results, [10] with 17,700 results and [12] with 391 results. The results were then ranked by relevance and manually filtered according the chosen criteria. In the end, not many works used GAN's in combination to Object Detection task as required. So 28 articles were chosen to be read and after that 14 were judged to have necessary relevance to this study. The results were organized in the table 1 considering the objective and proposed framework for each paper, and table 2 which shows the best Mean Average Precision (mAp) of the models when compared to the surrogate model used in each paper and the dataset used for evaluation:

Table 1. Selected Papers Proposed Frameworks

Method Objective	Paper	Proposed Framework
Overall Object Detection	[14]	SSD + GAN for Knowledge Distillation
	[15]	Competitive Object Detection Networks

Pedestrian Detection	[16]	GAN for synthetic data creation
	[17]	DCGAN + SSD
	[18]	DCGAN + SSD
	[19]	DCGAN + SSD
Small Object Detection	[20]	Faster R-CNN + GAN
	[21]	GAN + CNN + SSD/Faster R-CNN
	[22]	CNN + ResNet + GAN
Unsupervised Bounding Box Detection	[23]	CNN + GAN + Reinforcement Learning
	[24]	Dilated CNN + GAN with Mask Mean Loss
	[25]	Encoder + Conditional GAN

Table 2 - Best mAp(%) of models vs Surrogate Model

Paper	Best mAp(%) found when compared with Surrogate Model and Dataset	Surrogate Model	Dataset used for Evaluation
[14]	2.8	ResNet50	Pascal VOC 2007
[15]	2.56	SSD300	Pascal VOC 2007
[16]	Not Applicable	Not Applicable	Not Applicable
[17]	45.2	SSD	CIFAR-10/100
[18]	39.4	SSD	VOC
[19]	Not Applicable	Not Applicable	Not Applicable
[20]	19.47	Faster R-CNN	Tsinghua-Tencent 100K
[21]	25.1	FRCNN	COWC Dataset
[22]	60	Faster R-CNN (Small Objects)	Tsinghua-Tencent 100K
[23]	Not Applicable	Not Applicable	Not Applicable
[24]	5.37	[23]	Car (Stanford)
[25]	2.6	WCCN VGG16	VOC2007

The studied authors proposed many different approaches to address different tasks related to object detection, most of them with the goal of enhancing accuracy of traditional object detection frameworks.

For overall object detection purposes, [14] proposes a large trained network (teacher) for object detection and then uses a GAN for knowledge distillation from this network to a much simpler one (student) with better results in testing accuracy than the teacher model. The discriminator in this case check whether the results come from the teacher or the student nets and back propagate to the student network training. The results showed 2.8% better mAP than the surrogate studied.

[15] shows a baseline network trained with real data and competing with a generator net trained with augmented data, both network's results are then judged by a discriminator that distinguishes the results between the baseline and generator models. By the end of the adversarial training, the generator model can fool the discriminator and it is then used for inference and testing presenting 2.56% mAp better than the baseline model.

Synthetic data creation is a very popular application of GAN's as also shown by [9]. [16] uses GAN to generate real like pedestrian images from synthetic data generated by a game engine. This produces images of pedestrians in unusual scenarios and positions helping traditional object detection models to improve their ability to detect pedestrians.

[17], [18] and [19] use another very known and proven ability of Deep Convolutional GAN's which is images resolution improvement. All the three works combine the GAN's image resolution improvement with SSD to increase its pedestrian detection capabilities in different sizes and distances. This framework can be applied in many different scenarios and object classes with great improvements in mAP as shown in table 2.

Small Object Detection appears as an exponent application for GAN's. In [20] a GAN is embedded into a Faster R-CNN Network to generate residual representations of small objects to be similar to the ones of big objects which improves the detection ability of small objects when compared to a vanilla Faster R-CNN network. Results show around 19.5% better performance in small objects detection than a regular Faster R-CNN.

[21] created a framework that uses a GAN to create high resolution images from low resolution images as input, the discriminator compares a real high-resolution image to the generated image. It then uses a different CNN to detect edges and improve the resolution even further to finally use SSD or a Faster R-CNN to detect the objects, improving the mAP performance in 25.1%.

[22] approaches the small object detection task in a similar manner than [20], by using region proposals. The authors presented a perceptual GAN architecture where the generator creates super-resolved representations of small objects being supervised by the discriminator. The framework also uses a residual network in the generator to carry on the small object representation to be added in the last part of the generator and create super-resolved features. The discriminator also has a detection and classification branch to generate the bounding boxes and infer object class, showing 60% better mAP for small object detection than a Faster R-CNN.

One of the most exciting characteristics of GAN's is the ability to learn tasks in an unsupervised way, [23] and [24] uses this to identify bounding boxes. In [23], it is used to differentiate images generated by the network from the same images generated in a previous loop of the network in an iterative way. So, the discriminator output is used as reinforcement learning inputs to the model.

[24] approaches the problem in a different way. The generator creates a black mask while the discriminator compares the generated image to a ground truth. The training stops once the generator can fool the discriminator.

A novel ranking-discriminator network is proposed in [25] to verify the object class produced by a conditional GAN network trained with inputs an original image representation created by an encoder network. The framework uses image proposed regions to also identify the bounding box in a weakly supervised manner.

Unsupervised OD using GAN's is still a substantially unexplored field of work with many challenges to overcome. Although [23] and [24] proposes some novel approaches to this task, the results obtained are far from a state of the art OD framework and they also do not explore Multi Object Detection.

4. CONCLUSION

In this work, 14 papers were selected based on specific search criteria of object detection frameworks that utilizes GAN's in their methods. All the analyzed works proposed different approaches to the problem. Some of them attack the object detection as a generic problem while others have chosen to do specific tasks such as small objects or pedestrian detectors. Among the proposed frameworks GAN's were used as many types of tools: image enhancement, data generation, knowledge distillation.

In the end, although all the evaluated papers show very exciting results we can conclude that GAN's application to Object Detection task does not have a preferred framework to go to among the academic and Artificial Intelligence communities and therefore, present itself as very promising field of study to be developed in future works.

The relatively unexplored potential of unsupervised OD using GAN's was also something that was brought to attention in this work, and it is specifically exciting if we consider the many GAN frameworks that have been proposed with multiple applications and very interesting results.

5. REFERENCES

[1] ZOU, Zhengxia et al. Object detection in 20 years: A survey. **arXiv preprint arXiv:1905.05055**, 2019.

[2] GIRSHICK, Ross et al. Rich feature hierarchies for accurate object detection and semantic segmentation. In: **Proceedings of the IEEE conference on computer vision and pattern recognition**. 2014. p. 580-587.

[3] GIRSHICK, Ross. Fast r-cnn. In: **Proceedings of the IEEE international conference on computer vision**. 2015. p. 1440-1448.

[4] HE, Kaiming et al. Spatial pyramid pooling in deep convolutional networks for visual recognition. **IEEE transactions on pattern analysis and machine intelligence**, v. 37, n. 9, p. 1904-1916, 2015.

[5] REN, Shaoqing 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.

[6] REDMON, Joseph et al. You only look once: Unified, real-time object detection. In: **Proceedings of the IEEE conference on computer vision and pattern recognition**. 2016. p. 779-788.

[7] LIU, Wei et al. Ssd: Single shot multibox detector. In: **European conference on computer vision**. Springer, Cham, 2016. p. 21-37.

[8] GOODFELLOW, Ian et al. Generative adversarial nets. **Advances in neural information processing systems**, v. 27, 2014.

[9] NAVIDAN, Hojjat et al. Generative Adversarial Networks (GANs) in networking: A comprehensive survey & evaluation. **Computer Networks**, p. 108149, 2021.

[10] GOOGLE SCHOLAR. Initial Page. Available at: **<https://scholar.google.com/>**. Access in: Jul 25th, 2021

[11] SCIENCE DIRECT. Initial Page. Available at: **< https://www.sciencedirect.com/>**. Access in: Jul 25th, 2021

[12] IEEE Xplore. Initial Page. Available at: **<https://ieeexplore.ieee.org/Xplore/home.jsp >**. Access in: Jul 25th, 2021

[13] arXiv.og. Initial Page. Available at: **< https://arxiv.org/>**. Access in: Jul 25th, 2021

[14] WANG, Wanwei et al. Gan-knowledge distillation for one-stage object detection. **IEEE Access**, v. 8, p. 60719-60727, 2020.

[15] PRAKASH, Charan D.; KARAM, Lina J. It GAN DO better: GAN-based detection of objects on images with varying quality. **arXiv preprint arXiv:1912.01707**, 2019.

[16] HUANG, Shiyu; RAMANAN, Deva. Expecting the unexpected: Training detectors for unusual pedestrians with adversarial imposters. In: **Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition**. 2017. p. 2243-2252.

[17] DINAKARAN, Ranjith K. et al. Deep learning based pedestrian detection at distance in smart cities. In: **Proceedings of SAI Intelligent Systems Conference**. Springer, Cham, 2019. p. 588-593.

[18] DINAKARAN, Ranjith et al. Distant Pedestrian Detection in the Wild using Single Shot Detector with Deep Convolutional Generative Adversarial Networks. In: **2019 International Joint Conference on Neural Networks (IJCNN)**. IEEE, 2019. p. 1-7.

[19] DINAKARAN, Ranjith; ZHANG, Li; JIANG, Richard. In-vehicle object detection in the wild for driverless vehicles. In: **Developments of Artificial Intelligence Technologies in Computation and Robotics: Proceedings of the 14th International FLINS Conference (FLINS 2020)**. 2020. p. 1139-1147.

[20] HUANG, Wenqing; HUANG, Mingzhu; ZHANG, Yuting. Detection of traffic signs based on combination of GAN and faster-RCNN. In: **Journal of Physics: Conference Series**. IOP Publishing, 2018. p. 012159.

[21] RABBI, Jakaria et al. Small-object detection in remote sensing images with end-to-end edge-enhanced GAN and object detector network. **Remote Sensing**, v. 12, n. 9, p. 1432, 2020.

[22] LI, Jianan et al. Perceptual generative adversarial networks for small object detection. In: **Proceedings of the IEEE conference on computer vision and pattern recognition**. 2017. p. 1222-1230.

[23] HALICI, Eren; ALATAN, A. Aydin. Object Localization Without Bounding Box Information Using Generative Adversarial Reinforcement Learning. In: **2018 25th IEEE International Conference on Image Processing (ICIP)**. IEEE, 2018. p. 3728-3732.

[24] JANG, Heeoh et al. Generative object detection: Erasing the boundary via adversarial learning with mask. In: **2019 IEEE 2nd International Conference on Information Communication and Signal Processing (ICICSP)**. IEEE, 2019. p. 495-499.

[25] DIBA, Ali et al. Weakly supervised object discovery by generative adversarial & ranking networks. In: **Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops**. 2019. p. 0-0.