Article

Object Identification and Localization Using Grad-CAM++ with Mask Regional Convolution Neural Network

Xavier Alphonse Inbaraj ^{1,*}, Charlyn Villavicencio ^{1,2} , Julio Jerison Macrohon ¹ , Jyh-Horng Jeng ¹
and Jer-Guang Hsieh ³

¹ Department of Information Engineering, I-Shou University, Kaohsiung City 84001, Taiwan; charlyn.villavicencio@bulsu.edu.ph (C.V.); isu10903050D@cloud.isu.edu.tw (J.J.M.); jjeng@isu.edu.tw (J.-H.J.)

² College of Information and Communications Technology, Bulacan State University, Bulacan 3000, Philippines

³ Department of Electrical Engineering, I-Shou University, Kaohsiung City 84001, Taiwan; jghsieh@gmail.com

* Correspondence: xalphonse@gmail.com

Abstract: One of the fundamental advancements in the deployment of object detectors in real-time applications is to improve object recognition against obstruction, obscurity, and noises in images. In addition, object detection is a challenging task since it needs the correct detection of objects from images. Semantic segmentation and localization are an important module to recognizing an object in an image. The object localization method (Grad-CAM++) is mostly used by researchers for object localization, which uses the gradient with a convolution layer to build a localization map for important regions on the image. This paper proposes a method called Combined Grad-CAM++ with the Mask Regional Convolution Neural Network (GC-MRCNN) in order to detect objects in the image and also localization. The major advantage of proposed method is that they outperform all the counterpart methods in the domain and can also be used in unsupervised environments. The proposed detector based on GC-MRCNN provides a robust and feasible ability in detecting and classifying objects exist and their shapes in real time. It is found that the proposed method is able to perform highly effectively and efficiently in a wide range of images and provides higher resolution visual representation than existing methods (Grad-CAM, Grad-CAM++), which was proven by comparing various algorithms.

Keywords: Grad-CAM++; GC-MRCNN; Mask R-CNN; object localization; feature maps; object detection

Citation: Inbaraj, X.A.; Villavicencio, C.; Macrohon, J.J.; Jeng, J.-H.; Hsieh, J.-G. Object Identification and Localization Using Grad-CAM++ with Mask Regional Convolution Neural Network. *Electronics* **2021**, *10*, 1541. <https://doi.org/10.3390/electronics10131541>

Academic Editor:
Athanasios Voulodimos

Received: 28 May 2021
Accepted: 22 June 2021
Published: 25 June 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Copyright: © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

CNN-based algorithms have attained state-of-the-art performance for vision-based applications, classification, and object detection, thanks to their excellent feature learning capabilities. The interpretability of CNN's, on the other hand, is frequently questioned by the community, as these types of networks are complex to design and have many undefined parameters [1]. As a result, it is critical to create a CNN model that is simple to grasp and has all of the necessary parameters for real-time applications.

On the other hand, because deep layers in feature maps contain both semantic and spatial information, a fundamental difficulty in CNN-based learning is defining why classification CNN's predict what they predict. Grad-CAM [2] is one of the ways for locating certain areas of an input image so that CNN can identify the objects using feature maps. The Grad-CAM [2] algorithm requires a linear combination of feature maps, with each feature map's weight determined by the mean of its gradients. This concept is based on CAM [3], and Grad-CAM [2] is an extension of that method. Grad-CAM++ [4] is a further upgraded version of that algorithm.

It is known that Grad-CAM++ [4] is a very good algorithm among most well-known network interpretation algorithms. For sanity check-based tasks, Grad-CAM++ [4] yields perfect results among the recently studied state-of-the-art methods. Therefore, this study uses Grad-CAM++ as the one of the algorithms for object detection. Grad-CAM++ can

transfer directly to Occluding Patch as another interpretation algorithm. To obtain the results in an efficient way, this work uses a combination of Grad-CAM++ and Mask R-CNN [5].

Mask R-CNN is a two-step technique that first performs a coarse scan of the entire scenario before focusing on ROI [5,6]. After obtaining the confidences of object categories, the suggested model employs a sliding window-based method to determine bounding boxes straight from the top most feature map. As a result, the suggested method's (GC-MRCNN) framework is straightforward to construct and produces promising results in instance segmentation and object detection.

2. Motivations

Nowadays object detection has much more attention among researchers, many of the algorithms are derived for object detection and localization. Recently, the computer vision community has placed great efforts to gather the image data with different scenes, shapes, appearance, and various categories for data training. Furthermore, because data distribution biases deep learning systems, gathering data with uniform combinations of object attributes and faults is extremely difficult. As a result, it is critical to construct the model to address these challenges to achieve better object detection. In this research, the researchers used a very efficient approach called GC-MRCNN to improve object detection and localization in images.

3. Research Gap

Using conventional standards for object detection, deep learning techniques have already obtained state-of-the-art results. In 2016, Mask R-CNN outperformed the other detection models and won the COCO object detection competition. Due to the complex nature of the problem, the large number of annotated samples, and the wide range of object scales, Mask R-CNN's activities in object detection is hardly comparable.

4. Research Contribution

The key contribution of this paper is the use of Grad-CAM++ in combination with the Mask RCNN framework to recognize multi-scale objects in the COCO dataset with ResNet 101 as the backbone network. The proposed method effectively reduces detector box redundancy while increasing detection accuracy. The efficiency of the GC-MRCNN model is demonstrated by comparing various algorithms and results.

5. Related Works

Selvaraju, R.R. et al. [2] presents a class discriminative localization technique using Grad-CAM and also utilized visual explanations in order to make CNN-based models more transparent. Furthermore, to gain high-resolution visualization, Grad-CAM is combined with the existing high-resolution visualizations.

Zhou et al. [3] introduced a method called CAM for distinguishing locales utilized by a confined class of picture characterization. Conversely, it made existing best-in-class profound models interpretable without modifying their engineering, accordingly, evading the interpretability versus exactness compromise. The methodology is a speculation of CAM and is appropriate to an altogether more extensive scope of CNN model families: Discriminative locales utilized by a limited class of picture grouping. CNN's that do not contain any completely associated layers make existing cutting-edge profound models interpretable without adjusting their engineering, subsequently.

Cheng et al. [7] present a two-stage Faster RCNN-based approach, namely a Deep Adaptive Proposal Network (DAPNet). By feeding the input data into the backbone network, high level features are generated. The Category Prior Network (CPN) sub network and the Fine Region Proposal Network (F-RPN) used the aforementioned high-level features to obtain the category prior information and candidate regions for each image,

respectively. Finally, the accuracy detection sub network is put to the test for adaptive candidate box regression and classification using the accuracy detection sub network.

Ennadifi, E. et al. [8] proposed wheat disease detection using CNN. The training and testing dataset is gathered from CRA-W (Walloon Agricultural Research Center, Gembloux, Belgium), which has a total of 1163 images with two classes, such as affected and healthy. Then wheat spikes are segmented using Mask R-CNN from the image background and the disease is located using the GRAD CAM method, which highlights the required region from CNN model.

Alfred Schöttl et al. [9] presents CAM interpretability of a network with an updated loss function. In addition to that, the GRAD-CAM algorithm is utilized to evaluate interpretability measures with GradCAM entropy. The result section is explained by using ResNet 50 and PASCAL VOC data with efficient details.

Gupta, et al. [10] proposed a new approach for understanding different layers of neural networks and to identify the task that uses to break down models for analyzing applications and identifying objects. Work, probabilistic methodology, and angle-based methodologies have been utilized with the end goal of article restriction. Mathematical mean of heat maps of both the methodologies have been finished. In the previous methodology, the genuine article's slopes are made to stream into the last convolutional layer to decide the main focuses which would assist with foreseeing that specific item.

Wang, T. et al. [11] exhibited the preparation and testing of Mask R-CNN-based individual finders for top-see fisheye pictures. Indeed, even with the restricted assortment of comments on preparing information, this can fundamentally improve the location execution over the standard model. What is more, it has likewise researched two procedures appropriate for fisheye pictures and showed their capacities to additionally improve discovery results. Training with foundation pictures is additional material for frameworks with fixed cameras (fisheye or point of view).

Hoang, V.T. et al. [12] proposed a modified mask RCNN model which utilizes Mobile NetV3 as a backbone in the place of a residual network and to decrease the factors like FLOPs, size of the model, and time a depth wise separable convolution has been introduced.

In Wang, T. et al. [13], high-resolution photos are used to recognize the ship detection using Alex Net and generic feature extraction. The proposed technique surpasses You Only Look Once (YOLO) and SSD in terms of accuracy and IoU.

Nie et al. [14] also designed the model for ship detection by using the Mask RCNN algorithm. In addition to that, a soft non maximum suppression method is implemented to enhance the performance of the design.

Zhang, S. et al. [15] presents person detection using the RCNN algorithm which modified using aggregation loss. The presented algorithm enhances the similar objects to be close. To overcome image defect due to occlusion, a five-part pooling unit is built instead of a single layer.

Wang, X. et al. [16] proposed the repulsion loss as a bounding box regression loss. This loss causes an attraction to the target and a repulsion to other items in the environment. The repulsion prevents the object from relocating from its neighbors, which outperforms more crowd localization.

The Rotation Dense Feature Pyramid Network (R-DFPN) is a two-module approach described by Yang et al. [17]. For feature fusion, the first module employs a dense feature Pyramid Network (DFPN), while the second employs a Rotation Region Detection Network for prediction (RDN).

Cheng et al. [18] introduces the objective function, which is optimized using regularization constraint to train the rotation invariant layer. Furthermore, the network is fine-tuned by CNN to increase network performance. The author uses public NWPUVHR dataset for detection.

Without using guided back propagation methods, Mohammed Bany Muhammed et al. [19] produced the Class Activation Map. They obtain CAM, which is based on the output of convolutional layers but is independent of the class score. They also produced

the Class Activation Map on images without using any back propagation or changing the CNN architecture by using the first principle component of the combined weight matrix of the last convolution layer. Eigen-CAM is the name of the suggested approach.

Bae, W. et al. created a new benchmark dataset CUB200-2011, ImageNet-1K, and OpenImages30k to solve difficulties while constructing a class activation map using Global Average Pooling and extracting the bounding box by thresholding based on the class activation map's maximum value in [20].

Ye, W. et al. [21] proposed Probabilistic-CAM (PCAM) Pooling that perform the localization of CAM by a global pooling operation during training in probabilistic fashion that is only image-level supervision. They evaluated ResNet-34 as a backbone network. For a better localization heatmap, they used ResNet-34 with LSE pooling with a 0.9-threshold value. By applying a 0.9-threshold value, they generate the bounding boxes that isolated regions on binary masks.

Xia, S. et al. [22] introduced the WS-AM method that combined two methodologies, the Grad-CAM and Weakly Supervised information. By using this method, the author extracts the regions that contains category-specific objects in an image. They used different pretrained networks (ImageNet-CNN and Places365-CNN) for feature extraction in SoftMax.

In [23], Bae, W. et al. examined some flaws during employed Class Activation Mapping (CAM). They provide some solutions to higher weights of gradient in activation area, negatively weights in the object regions and instability due to maximum threshold value. Further, they proposed Threshold Average Pooling (TAP) to alleviate the problem of Global Average Pooling (GAP).

By not employing Global Average Pooling and Group Normalization in models, Preechakul, K. et al. [24] suggested the Pyramid Localization Network (PYLON) to achieve better accuracy than a typical CAM model such that its average points are higher than 0.17 on the NIH's Chest X-ray 14 dataset.

Lin, Y. et al. used a VGG 16 backbone encoder and decoder architecture with skip connections method in [25]. They introduced a defect segmentation network with a feedback refinement mechanism, commonly known as CAM-UNet. There are two major steps to it. One is referred to as pretraining, while the other is referred to as feedback refining. During the pretraining phases, a Class Activation Map (CAM) is created, which is then adjusted to the weights of the pretrained models. Using the Class Activation Map in the feedback refinement processes, the segmentation output is gradually improved.

6. Proposed Method

In proposed method, initially Grad-CAM++ [4] model was designed to localize the object and to generate the visual explanation for a RCNN network without modifying its structure. Then Mask R-CNN [5] was utilized to detect the object which is located by the Grad-CAM++ [4].

The background on Grad-CAM++ [4] visualization considers a deep network for object recognition. The output of Grad-CAM++ [4] is a heatmap of hot regions of the specific class.

The localization map or heat map is obtained by, $L_{Grad-CAM++}^c \in \mathbb{R}^{u \times v}$, where u is the width of the class object, v is the height of the class object, and c is the class.

Let us calculate the gradient score for class c and y^c with respect to A^k of the convolution layer ($\frac{y^c}{A^k}$).

Then the weights α_k^c is defined and propagated by backpropagation and through global average pooling by Equation (1) and by linear combination by Equation (2):

$$\alpha_k^c = \frac{1}{Z} \sum_i \sum_j \frac{\partial y^c}{\partial A_{ij}^k} \quad (1)$$

$$L^c = \text{ReLU}\left(\sum_k \alpha_k^c A^k\right) \quad (2)$$

where A^k is the feature map activation and y^c is the neural network output before performing softmax.

Hence, α_k^c represents a partial linearization of the deep network downstream from A which evaluates the ‘importance’ for feature map k with class c .

Then the Grad-CAM++ [4] heat map calculates the weighted combination of feature map activation A^k with weights α_k^c .

For each feature map, α is the corresponding weight and computes the weighted sum of A^k which is required output. It is noted that the convolutional feature maps have the same size of a heat map (14×14) in the case VGG [26] and AlexNet [27] networks. Then to separate the positive value ReLU operation is applied, so that the negative values will be zero. In a case without ReLU, localization will be worse since it highlights more than just the desired class.

The general connection between GradCAM and Class Activation Mapping (CAM) is discussed in [3], and it has been proved that for a various CNN structures Grad-CAM generalizes CAM.

CAM [3] generates a heat map for classification with a structure of CNN where global average pooled convolutional feature maps are fed directly into softmax.

Let us consider that the Global Average Pooling (GAP) is used to pool the feature map and is transformed linearly to compute score Y^c for each class c , then,

$$Y^c = \sum_k w_k^c \frac{1}{Z} \sum_i \sum_j A_{ij}^k \quad (3)$$

where Z is the number of pixels in the feature map.

Let F^k be the global average pooled output,

$$F^k = \frac{1}{Z} \sum_i \sum_j A_{ij}^k. \quad (4)$$

Then the score is:

$$Y^c = \sum_k w_k^c \cdot F^k \quad (5)$$

where w_k^c is the weight of class c with feature map k . Apply the gradient of the score for class c (Y^c) with respect to F^k ,

$$\frac{\partial Y^c}{\partial F^k} = \frac{\frac{\partial Y^c}{\partial A_{ij}^k}}{\frac{\partial F^k}{\partial A_{ij}^k}}. \quad (6)$$

Taking partial derivative of Equation (4):

$$\frac{F^k}{A_{ij}^k} = \frac{1}{Z}. \quad (7)$$

Substituting this in Equation (6), derives:

$$\frac{\partial Y^c}{\partial F^k} = \frac{\partial Y^c}{\partial A_{ij}^k} \cdot Z. \quad (8)$$

From Equation (5) derives that,

$$\frac{\partial Y^c}{\partial F^k} = w_k^c. \quad (9)$$

Hence,

$$w_k^c = Z \cdot \frac{\partial Y^c}{\partial A_{ij}^k}. \quad (10)$$

Summing both sides of Equation (10) over all pixels (i, j) ,

$$\sum_i \sum_j w_k^c = \sum_i \sum_j Z \cdot \frac{\partial Y^c}{\partial A_{ij}^k}. \quad (11)$$

Since Z and w_k^c do not depend on i and j , rewriting this as:

$$Z w_k^c = Z \sum_i \sum_j \frac{Y^c}{A_{ij}^k}. \quad (12)$$

Note that Z is the number of pixels in the feature map. Thus, when the terms are re-ordered and see that:

$$w_k^c = \sum_i \sum_j \alpha_{ij}^{kc} \cdot \text{ReLU} \left(\frac{\partial Y^c}{\partial A_{ij}^k} \right). \quad (13)$$

The equation for w_k^c is identical to the α_k^c used by Grad-CAM++ up to a proportionality constant $(1/Z)$ that gets normalized out during visualization.

6.1. Mask R-CNN

It starts with the process of sorting and refining the anchors. A positive and negative anchor is proposed around every object which is then refined further. These anchors can be customized according to the kind of objects one wants to detect. Furthermore, these anchors play a significant role in the overall performance of Mask R-CNN [5].

This is followed by the appearance of bounding boxes or regions as a result of the Region Proposal Network (RPN). This is a very crucial point because the Region Proposal Network decides whether a particular object is a background or not and likewise bounding boxes are marked. The objects in the frame are determined by a classifier and a regressor by region of interest pooling. After refining the bounding boxes, masks are generated and placed on their appropriate positions on the object. This is the main distinguishable attribute of Mask R-CNN [5].

6.2. Backbone Network

This is the first network that processes the image in Mask R-CNN [28]. Generally, ResNet 50 or ResNet 101 is used for feature extraction. Primary features are extracted by the initial layers of this network. Whereas high-level feature extraction is accomplished by proceeding layers. The objects of different scales are detected by the Feature Pyramid Network (FPN). Even though pyramid rendering is eschewed because of computational and memory demand, a pyramid hierarchy is used in order to obtain feature pyramids without heavily compromising the efficiency. FPN allows access of features at higher as well as lower levels. Objects in images will be passed in various scales that will be detected to determine features by FPN that are the main backbone of this proposed method, shown in Figure 1.

It can be termed as the backbone of Mask R-CNN [28]. FPN is the first phase in Faster R-CNN. During this stage, the most suitable bounding boxes are determined along with the objectness score. These anchor boxes might vary in size because of the variance of object dimensions in the image. The classifier is trained in order to distinguish the background and foreground. Moreover, if the anchors are turned to vectors with two values and then fed to an activation function, labels can be predicted. A regressor of bounding box refines

the anchors and a regressor loss function determines the regression loss by following Equations (14) and (15):

$$L_{loc}(t^u, v) = \sum_{i \in \{x, y, w, h\}} \text{smooth}_{L_1}(t_i^u - v_i), \quad (14)$$

$$\text{smooth}_{L_1}(x) = \begin{cases} 0.5x^2 & \text{if } |x| < 1 \\ |x| - 0.5 & \text{otherwise} \end{cases} \quad (\text{smooth}_{L_1} \text{ Loss Function of the Regressor}) \quad (15)$$

The region proposal network determines the most appropriate bounding boxes for the objects with the help of classification loss and regression loss.

Figure 1. Backbone network of Mask R-CNN [28].

6.3. ROI Pooling and Classifier Training

The RPN gives the most appropriate regions of bounding boxes. The predicament here is that they are of different dimensions. To eradicate this difference, ROI Pooling is applied in order to commensurate the feature maps of CNN [12]. In other words, ROI Pooling makes the process easier and efficient by converting many feature maps of different sizes into the same dimensions which makes a feasible structure for further processing.

The feature map is divided into a fixed number of equal sized regions and then max pooling is applied to them. This makes the output of ROI Pooling constant irrespective of the size of the input. In this stage, features are extracted and can also be used for further conjectures. After this, RPN, the classifier, and the regressor can be trained altogether as well as distinctively. When the aforementioned three are trained together, the speed spikes up to 150% while the accuracy remains unhindered. Stochastic gradient descent and back-propagation can be used to train the region proposal network by following the image-centric sampling. Figure 2 is mainly focused on detecting objects after performing ROI pooling and Classifier Training.

Mask Generation and Layer Activation Mask is generated upon objects simultaneously while bounding boxes are predicted by the FPN. The process of instance segmentation is the key feature behind generating the masks for the objects in an image. This occurs at the pixel level and is the extension of a Faster R-CNN which results in a Mask R-CNN.

Low-intensity masks are generated in accordance with the classifier. These soft masks are more meticulous than binary masks. The loss is calculated during training and then the object mask is augmented to the size of bounding boxes. Moreover, activations of layers can be examined individually for detecting the noises (if any) and to rectify it. This generates a mask for each object in accordance with the input.

Figure 2. Mask R-CNN for object detection [19].

7. Results and Discussion

This work is carried out using MATLAB 2020a and uses ResNet 101 as the backbone network and COCO to evaluate. To pre train the dataset, Mask R-CNN [5] is utilized. The images in the dataset are resized into 256×256 before the training process. To ease the training process of Mask R-CNN for larger data the learning rate is chosen as 0.00001, While in the network training, the Grad-CAM++ [20] algorithm is used to update the weight function, the initial layers are evaluated for first two epochs, and then the remaining layers are trained with Mask R-CNN [5] along with learning rate annealing. After testing several groups of parameter combinations empirically the best output is obtained using learning rate annealing starting with 0.00001. Once the class labels are correctly distinguished then Grad-CAM++ visualization can localize the correct object. Finally, to remove the detection result of a similar class, post processing is applied over union among bounding boxes. This will maintain the bounding boxes with a larger area and removing the bounding boxes with a smaller area.

If the class label is matched with the ground truth, then the corresponding bounding boxes will be chosen. Examining the proposed method by passing various input images to predict the proper object, the first input images were classified as a feline tiger dog as shown in Figure 3.

Figure 3. (a) Input image; (b) output of Grad-CAM++ shows that the Grad-CAM++ representations for ‘feline tiger dog’; and (c) detected dog using Mask R-CNN shows that the recognized canine utilizing the proposed system (GC-MRCNN).

Figure 4 shows that results of various input images and its outputs results of Grad-CAM, Grad-CAM++, and the proposed algorithm (GC-MRCNN): (a) Original image, (b) Class Activation Map, (c) results of Grad-CAM, (d) results of Grad-CAM++, and (e) GC-MRCNN. Here class labels did not predict because the researchers used cat and dog categories of images only from the COCO dataset. However, in Figure 5, you can observe the class label with the proposed algorithm.

Figure 4. (a) Original image; (b) Class Activation Map; (c) results of Grad-CAM; (d) results of Grad-CAM++; and (e) GC-MRCNN. Various original images and its results of Grad-CAM, Grad-CAM++, and the proposed algorithm (GC-MRCNN).

Figure 5 demonstrates the detected object using the proposed GC-MRCNN. It can be observed that the proposed algorithm localizes and detects the object perfectly while comparing with other algorithms: (a) Original image, (b) Class Activation Map, (c) Grad-CAM, (d) Grad-CAM++, and (e) GC-MRCNN (proposed).

Figure 6 shows the difference between using a Mask R-CNN and not using a Mask R-CNN. We can observe that the regression box on object is not predicted properly in the image, which is a major difference: (a) Original Image, (b) results of Class Activation Mapping, (c) results of Grad-CAM, (d) results of Grad-CAM++, (e) results with Mask R-CNN, and (f) results without Mask R-CNN.

Figure 5. (a) Original image; (b) Class Activation Map; (c) Grad-CAM; (d) Grad-CAM++; and (e) GC-MRCNN (proposed). The figure demonstrates the detected object in the image using a proposed algorithm (GC-MRCNN).

Figure 6. (a) Original Image; (b) results of Class Activation Mapping; (c) results of Grad-CAM; (d) results of Grad-CAM++; (e) results with Mask R-CNN; and (f) results without Mask R-CNN. The results of difference between with R-CNN and Without R-CNN.

Table 1 shows a comparison chart of the detection rate of Grad-CAM [2] and Grad-CAM++. The researchers analyzed and identified the detection rate of Grad-CAM++ for the proposed algorithm. The researchers did not mention any citation on Grad-CAM++. The detection rate is category-specific visualization techniques. When viewing Grad-CAM [2], it can correctly identify the categories being visualized in 61.23% of cases. However, compared to Grad-CAM [2], Grad-CAM++ showed improved performance (17.02%) in correctly identify category-specific visualization.

Table 1. Detection rate comparison.

Method	Grad-CAM [2]	Grad-CAM++
Detection Rate	61.23	78.25

Table 2 shows the Intersection over Union (IoU) of Grad-CAM [2] and Grad-CAM++. It can be observed that the researchers did not cite on Grad-CAM++ because of the threshold value that the researchers detected during the process of the proposed algorithm on localizing object in an image. In the proposed algorithm (GC-MRCNN), Grad-CAM++ is mainly used to localizing the feature map. However, detecting the object in that feature map was performed by Mask-RCNN. Thus, the researchers updated the weights of GradCAM++ into a Mask R-CNN. Finally, proper detection is achieved on an object in an image. Finding detection rate (Table 1) and Intersection over Union (IoU) (Table 2) are mainly used to analyze and detect the localization on an image. Based on that, the researchers incorporated the weights of a feature map in detecting the object in an image then showed in the proposed GC-MRCNN (Table 3).

Table 2. Intersection over Union (IoU) results for object localization.

Threshold Value	0	0.25	0.50
Grad-CAM [2]	0.33	0.28	0.16
Grad-CAM++	0.339	0.3	0.2

Table 3. Average over IoU Thresholds (AP) for different algorithms.

Algorithms	AP	AP ₅₀	AP ₇₅
MNC [29]	24.6	44.3	24.8
FCIS+OHEM [30]	29.2	49.5	NA
(FCIS++)+OHEM [30]	33.6	54.5	NA
Mask R-CNN [5]	33.1	54.9	34.8
GC-MRCNN (Proposed)	49.57	71.88	65.25

Table 3 shows the Average over IoU Thresholds (AP) of several algorithms, as well as a comparison to the proposed GC-MRCNN. This table illustrates that the suggested algorithm outperforms other algorithms in terms of Average over IoU Thresholds (AP). This Average over IoU Thresholds (AP) will prove to be valuable in localizing and recognizing each item instance within an image. Localize each object instance using bounding boxes, then split each object by categorizing each pixel into a set of preset categories. The researchers compared some scale measures with the proposed GC-MRCNN, which are AP, AP₅₀, and AP₇₅. IoU is a useful metric for determining how much overlap exists between two bounding boxes or segmentation masks. $\text{IoU} = 1$ if the prediction is perfect, and $\text{IoU} = 0$ if it entirely fails. $\text{AP}@ \alpha$ or $\text{AP} \alpha$ stands for Average over at IoU threshold of α . As a result, AP, AP₅₀, and AP₇₅ denote AP at IoU thresholds of 0.5:0.95 (from 0.5 to 0.95, step 0.05 (0.5, 0.55, 0.6, 0.65, 0.7, 0.75, 0.8, 0.85, 0.9, and 0.95), 50% and 75%, respectively. A

threshold is used to objectively determine whether the model successfully predicted the box placement. There is considerable overlap between the expected box and one of the ground-truth boxes if the model predicts a box with an IoU score larger than or equal to the threshold. This indicates that the model was effective in detecting an object. Positive is the classification for the detected region (i.e., contains an object).

When the IoU score is less than the threshold, on the other hand, the model generated a faulty prediction since the projected box does not overlap with the ground-truth box. This indicates that the discovered region has been labeled as negative (i.e., does not contain an object). The researchers used the term “Average over IoU Threshold (AP)” as in [5], however it is mean Average Precision (mAP) in other comparisons [29,30]. The proposed GC-MRCNN method outperformed other algorithms in terms of predicting box on object.

Figure 7 shows that the result of predicting object accuracy is 82% in the proposed algorithm (GC-MRCNN). In the proposed GC-MRCNN, the researchers used ResNet 101 as the backbone network and the COCO dataset for evaluation. COCO is a large-scale dataset that includes object detection, segmentation, key-point detection, and captioning. There are 328 K photos in the dataset.

Figure 7. Result of predicting object accuracy of the proposed algorithm (GC-MRCNN).

In 2014, the first version of the COCO dataset was released. It has 164 K photos divided into three sets: Training (83 K), validation (41 K), and test (41 K). A fresh test set of 81 K photographs was released in 2015, which included all the previous test images as well as 40 K new images.

The GC-MRCNN visualizations of the model predictions revealed that it had learned to provide better detection on images. According to Equation (1), evaluating the ‘importance’ for feature map k with a class c can act as a concept of ‘detector’. If it has higher positive values, then it leads to the presence of detection, otherwise it leads to an absence of class score. Thus, leading in positive class score was shown by an accuracy of 82%.

8. Conclusions

This work proposed a GC-MRCNN algorithm to localize and detect the objects for any CNN network model by giving visualization. Moreover, Grad-CAM++ localization produces a high-resolution output in comparison with existing methods and also produced the best texture-based model decision to localize the object easily. GC-MRCNN was used for localizing an object by region proposal that is based on producing a heatmap in an image that will evaluate more accurately through predicting a regression box for the object. This work can be extended by including decision-based rules using deep networks in fields like video applications, reinforcement learning, and natural language processing.

Author Contributions: Conceptualization, X.A.I.; methodology, C.V. and J.J.M.; validation, J.J.M. and C.V.; formal analysis, C.V. and J.J.M.; writing—original draft preparation, X.A.I.; writing—review and editing, J.J.M., X.A.I., C.V., and J.-H.J.; visualization, C.V.; supervision, J.-H.J. and J.-G.H.; project administration, J.-H.J. and J.-G.H. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded as a scholar of the Ministry of Science and Technology (MOST), Taiwan and I-Shou University, Kaohsiung City, Taiwan.

Data Availability Statement: <https://cocodataset.org/#download> (accessed on 28 May 2021).

Acknowledgments: The authors wish to thank I-Shou University, Taiwan. This work was supported in part as a scholar of Taiwan's Ministry of Science and Technology (MOST). The researchers would also like to thank the Almighty God for His guidance from the start until the completion of this study.

Conflicts of Interest: The authors declare no conflict of interest.

References

- Li, J.; Schmidt, F.; Kolter, Z. Adversarial camera stickers: A physical-camera-based attack on deep learning systems. In Proceedings of the International Conference on Machine Learning, Long Beach, CA, USA, 10–15 June 2019; pp. 3896–3904.
- Selvaraju, R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; Batra, D. Grad-CAM: Visual Explanations From Deep Networks via Gradient-Based Localization. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017.
- Zhou, B.; Khosla, A.; Lapedriza, A.; Oliva, A.; Torralba, A. Learning Deep Features for Discriminative Localization. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016.
- Chattopadhyay, A.; Sarkar, A.; Howlader, P.; Balasubramanian, V.N. Grad-CAM++: Improved Visual Explanations for Deep Convolutional Networks. *arXiv* **2018**, arXiv:1710.11063v3.
- He, K.; Gkioxari, G.; Dollar, P.; Girshick, R. MASK RCNN. *arXiv* **2018**, arXiv:1703.06870v3.
- Sermanet, P.; Eigen, D.; Zhang, X.; Mathieu, M.; Fergus, R.; LeCun, Y. OverFeat: Integrated Recognition, Localization and Detection using Convolutional Networks. *arXiv* **2014**, arXiv:1312.6229.
- Cheng, L.; Liu, X.; Li, L.; Jiao, L.; Tang, X. Deep Adaptive Proposal Network for Object Detection in Optical Remote Sensing Images. *arXiv* **2018**, arXiv:1807.07327.
- Ennadifi, E.; Laraba, S.; Vincke, D.; Mercatoris, B.; Gosselin, B. Wheat Diseases Classification and Localization Utilizing system network. In Proceedings of the 2020 International Conference on Intelligent Systems and Computer Vision (ISCV), Fez, Morocco, 1–5 June 2020.
- Schöttl, A. A light-weight method to foster the (Grad)CAM interpretability and explainability of classification networks. In Proceedings of the 2020 10th International Conference on Advanced Computer Information Technologies Alfred Schöttl, Deggendorf, Germany, 13–15 May 2020.
- Gupta, S.; Bagga, S.; Dharandher, S.K.; Sharma, D.K. GPOL: Gradient and Probabilistic approach for Object Localization to understand the working of CNNs. In Proceedings of the 2019 IEEE Bombay Section Signature Conference (IBSSC), Mumbai, India, 26–28 July 2019.
- Wang, T.; Hsieh, Y.Y.; Wong, F.W.; Chen, Y.F. Mask-RCNN Based People Detection Using A Top-View Fisheye Camera. In Proceedings of the 24th International Conference on Technologies and Applications of Artificial Intelligence, TAAI 2019, Kaohsiung, Taiwan, 21–23 November 2019.
- Hoang, V.-T.; Hoang, V.-D.; Jo, K.-H. Realtime Multi-Person Pose Estimation with RCNN and Depthwise Separable Convolution. In Proceedings of the 2020 RIVF International Conference on Computing and Communication Technologies (RIVF), Ho Chi Minh City, Vietnam, 6–7 April 2020.
- Wang, T.; Gu, Y. Cnn Based Renormalization Method for Ship Detection in Vhr Remote Sensing Images. In Proceedings of the IGARSS IEEE International Geoscience and Remote Sensing Symposium, Valencia, Spain, 22–27 July 2018; pp. 1252–1255.
- Nie, S.; Jiang, Z.; Zhang, H.; Cai, B.; Yao, Y. Inshore Ship Detection Based on Mask R-CNN. In Proceedings of the IGARSS IEEE International Geoscience and Remote Sensing Symposium, Valencia, Spain, 22–27 July 2018; pp. 693–696.
- Zhang, S.; Wen, L.; Bian, X.; Lei, Z.; Li, S.Z. Occlusion-aware r-cnn: Detecting pedestrians in a crowd. In Proceedings of the European Conference on Computer Vision, Munich, Germany, 8–14 September 2018; pp. 637–653.
- Wang, X.; Xiao, T.; Jiang, Y.; Shao, S.; Sun, J.; Shen, C. Repulsion loss: Detecting pedestrians in a crowd. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 July 2018; pp. 7774–7783.
- Yang, X.; Sun, H.; Fu, K.; Yang, J.; Sun, X.; Yan, M.; Guo, Z. Automatic Ship Detection in Remote Sensing Images from Google Earth of Complex Scenes Based on Multiscale Rotation Dense Feature Pyramid Networks. *Remote. Sens.* **2018**, *10*, 132. [[CrossRef](#)]
- Cheng, G.; Zhou, P.; Han, J. Learning Rotation-Invariant Convolutional Neural Networks for Object Detection in VHR Optical Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2016**, *54*, 7405–7415. [[CrossRef](#)]
- Muhammad, M.B.; Yeasin, M. Eigen-CAM: Class Activation Map using Principal Components. In Proceedings of the 2020 International Joint Conference on Neural Networks (IJCNN), Glasgow, UK, 19–24 July 2020. [[CrossRef](#)]

20. Bae, W.; Noh, J.; Kim, G. Rethinking Class Activation Mapping for Weakly Supervised Object Localization. In Proceedings of the CVPR 2020, Virtual, Seattle, WA, USA, 14–19 June 2020.
21. Ye, W.; Yao, J.; Xue, H.; Li, Y. Weakly Supervised Lesion Localization With Probabilistic-CAM Pooling. *arXiv* **2020**, arXiv:2005.14480.
22. Xia, S.; Zeng, J.; Leng, L.; Fu, X. WS-AM: Weakly Supervised Attention Map for Scene Recognition. *Electronics* **2019**, *8*, 1072. [[CrossRef](#)]
23. Bae, W.; Noh, J.; Kim, J.S.G. Revisiting Class Activation Mapping for Learning from Imperfect Data. In Proceedings of the CVPR 2020, Virtual, Seattle, WA, USA, 14–19 June 2020.
24. Preechakul, K.; Sriswasdi, S.; Kijssirikul, B.; Chuangsuwanich, E. High resolution weakly supervised localization architectures for medical images, ICASSP. *arXiv* **2021**, arXiv:2010.11475.
25. Lin, D.; Li, Y.; Prasad, S.; Nwe, T.L.; Dong, S.; Oo, Z.M. CAM-UNET: Class Activation MAP Guided UNET with Feedback Refinement for Defect Segmentation. In Proceedings of the 2020 IEEE International Conference on Image Processing (ICIP), Abu Dhabi, United Arab Emirates, 25–28 October 2020; pp. 2131–2135. [[CrossRef](#)]
26. Simonyan, K.; Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv* **2014**, arXiv:1409.1556.
27. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. *Adv. Neural Inf. Process. Syst.* **2012**, *25*, 1097–1105. [[CrossRef](#)]
28. Mishra, S.; Tare, P. Computational Calibres of Mask R-CNN in Real-Time Object Detection and Instance Segmentation. *Int. J. Innov. Res. Sci. Eng. Technol.* **2019**, *8*, 1744–1750.
29. Dai, J.; He, K.; Sun, J. Instance-aware semantic segmentation via multi-task network cascades. In Proceedings of the CVPR 2016, Las Vegas, NV, USA, 27–30 June 2016.
30. Li, Y.; Qi, H.; Dai, J.; Ji, X.; Wei, Y. Fully convolutional instance-aware semantic segmentation. In Proceedings of the CVPR 2017, Honolulu, HI, USA, 21–26 July 2017.