Course Project Report

One-Shot Object Detection with Co-Attention andCo-Excitation

Submitted By

Jaidev Chittoria (181IT119)
Ayush Bhandari (181IT209)

as part of the requirements of the course

Computer Vision (IT465) [Jul - Nov 2021]

in partial fulfillment of the requirements for the award of the degree of

Bachelor of Technology in Information Technology

under the guidance of

Mr. Dinesh Naik , Dept of IT, NITK Surathkal

undergone at

DEPARTMENT OF INFORMATION TECHNOLOGY

NATIONAL INSTITUTE OF TECHNOLOGY KARNATAKA, SURATHKAL

JUL-NOV 2021

One-Shot Object Detection with Co-Attention and Co-Excitation

Jaidev Chittoria[1], Ayush Bhandari [2]

***Abstract*—This paper tries to aim to solve challenges regarding one shot object detection.Basically on providing a query image patch whose class label is not provided in the training, the aim is to detect all the occurences of the same class in the target image and putting bounding box on the occurences . To do so co-attention and co-excitation (CoAE)[1] framework will be used which will do work in the three step process staring from using the non local operation to inspect the each query-target pairs leading to getting region of importance . Then formulating a it such that it can adaptively emphasize on correlated feature channels which will lead to identifying possible target objects. And at last going through each target region and apply a margin based ranking which basically predicts the similarity between possible target region and querying object respective of its class label [2].**

I. INTRODUCTION

Humans have the tendency to learn new concepts with some guidance and then to apply the concepts in real world eg. learning to identify and find objects in image based on the given small piece of information whether it written for in image format.Humans can even classify objects without knowing any about how data can be classified and processed such as counting and grouping the pixels of object as one and extracting vaious information for comparsion ,etc and all these thing can be done on a wide variation in appearances, viewing angles and lighting etc. But if we try to replicate these things on a machine it becomes quite tedious and challenging and even if the models are built they don't perform well on various performance metrics so we would try to build a system that can perform one shot object detection with good results [4].

A. Motivation

One shot object detection is a categorization problem and in general in machine learning to do this it requires a training on a very large dataset i.e high computational power to learn about object categories, identify the number of occurences of object in image under various conidition ranging from lighting ,shade ,aspect ratio etc and in one shot it tries to learn information from one or few images of object and to find them on target images. And to do so one must have proper knowledge about how to classify the objects? ,What are widely used models? ,What function to use to compare the features between images and how to experiments with different models? What all performance metrics should be consider for evaluation of model?,etc.

II. LITERATURE SURVEY

A. Related Work

1) Leveraging Bottom-Up and Top-Down Attention for Few-Shot Object Detection: To improve the performance and interpretability of few-shot object detectors, the author has proposed an attentive few-shot object detection network (AttFDNet) that takes the advantages of both top-down and bottom-up attention

2) Few-Shot Object Detection with Attention-RPN and Multi-Relation Detector: The author presented a few-shot object detection which aims to detect objects of unseen class with a few training examples and a new dataset was created with 1000 categories of object and the model also had a wide range of applicability [2,5].

3) Comparison Network for One-Shot Conditional Object Detection: The one-shot detection has been presented as a conditional probability problem. Working towards this, a novel one-shot conditional object detection (OSCD) framework, referred as Comparison Network (ComparisonNet), has been presented. Experiments show that the proposed approach achieves state-of-the-art performance on the proposed datasets of Fashion-MNIST and PASCAL VOC [3,7].

4) Siamese Neural Networks for One-shot Image Recognition: Proposed the method for learning siamese neural networks which use a special structure which helps in ranking similarity between inputs.The method was able to outperforms all baselines by good margin but they only incur the dataset which involves alphabets in images [8].

5) Transfer Learning by Borrowing Examples for Multiclass Object Detection: Paper discusses how to overcome the lack of training data for certain classes and also provides a novel way of augmenting the training data for each class by borrowing and transforming examples from other classes. It also a improvement over the state of art detector on dataset [5,9].

B. Outcome of Literature Review

After referring papers we were able to understand the challenges related to one shot object detection as well as importance of feature vector for query and target image and the importance of similarity score and the importance of the evaluation metrics for model. Also authors have the use ResNet50 and siamese network so we would like to experiment with different neural networks to understand which fits the problem well and to analyse the results on different network.

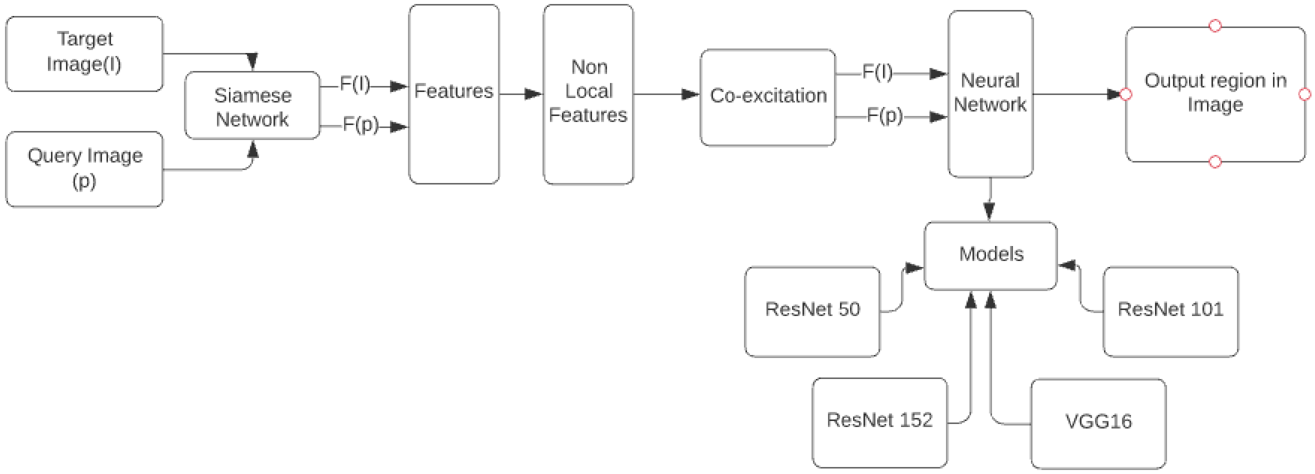

Fig. 1: Architecture Design

C. Problem Statement

Solving the problem of one shot object detection i.e when a similar one shot task of localization and perceptual categorization is performed using a framework (co-attention and co-excitation).

D. Objective

1) Data Preprocessing and annotation of Images.
2) Developing siamese network to generate feature maps with respect to target-query image pair for similarity calculation.
3) Performing co attention and co excitation where a local operation is applied to feature maps with respect to query and target image in which the element wise sum over the feature maps is applied.
4) Final inference to predict the bounding box location of the query image in the target image.

III. Methodology

A. Procedure

1) First we have query-target image pair where the query image contains the object image which we tend to detect in the target image by finding the bounding box relative pixel coordinates. These two input images will be fed to siamese network [10] which further will extract the feature maps [11] of both images which are similar.
2) Non local object feature sproposal will be used which will lead to lesser and more accurate number of regions for one shot learning.
3) On the feature maps of query and target image reweighting will be done which will be distributed over N channels. It will lead to spatially summary of each feature map with global average pooling and also it will act as bridge between query and target image to work on important feature channels which will help in finding the similarity between the images. Further a non local operation will be applied to these feature maps which will do the element wise sum pair of the feature maps. These two steps are coattention and coexcitation [12,13,14].
4) After that the feature maps will be fed to the final neural network which will use a marginal loss based score to generate bounding box location of the query object in the target image. Finally using a bounding box threshold value it will be decided whether the prediction is correct or not [15,16].

B. Experimentation with models

All of the following listed models are varying in terms of layers as well as performance but all of them have pretrained model which is trained on million of ImageNet dataset.

1) ResNet50 :It is a CNN that is having 50 layers.
2) ResNet101 :It is a CNN that is having 101 layers.
3) ResNet152 :It is a CNN that is having 152 layers.
4) vgg16 :It is a CNN that is having 16 layers.

nd also pretrained version of the network can be used which is trained on over million images from the ImageNet.

C. Hyperparameter Tuning

Since hyperparameters plays an important role as it define behaviour of model and so finding a combination of hyperparameters which basically gives optimum result and also helps in minimizing the value of predefined loss function.With few experimentation we were able to observe the important parameters for the models and those were Learning rate, Gamma, Threshold and Batch size [17].

D. Performance Metrics

To evaluate the performance of the model the following metrics has been considered

1) IoU
2) Average Precision
3) Average Recall
4) Mean Average Precision

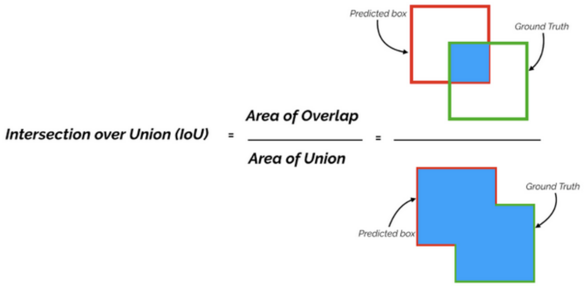

Fig. 2: IOU

IV. Result and Analysis

The image dataset considered for this project is Imagenet dataset we have taken images in which 118,000 images were taken for training, 41,000 images were taken for testing and 5,000 images were taken for validation. For annotations purpose we have used COCO(common object in context) api [18,19].

S.No	Objects	Object id
1	Cake	56
2	Chair	57
3	Bed	60
4	Book	74
5	Clock	75
6	Tv	63
7	Teddy Bear	78
8	Mouse	65
9	Keyboard	67
10	Oven	78

TABLE I: Some of classes of objects

The number of classes taken into consideration can be seen in figure.3. There are about 80 classes like person, bicycle, car, motorcycle, banana, cow, sheep etc. These classes are the required objects that the model intend to find in the target image.

S.No	Hyperparameters	Value
1	Batch Size	128
2	Learning rate	0.001
3	Threshold	0.5
4	Gamma	0.1

TABLE II: Hyperparameters of neural networks

In table 2, on experimentation we observe that these hyperparameters were playing a significant role for the neural network model in the above mentioned range of value.

Iteration	IOU	Area	Average Precision	Average Recall
1	0.50:0.95	all	0.116	0.129
2	0.50	all	0.231	0.260
3	0.75	all	0.109	0.292
4	0.50:0.95	small	0.052	0.154
5	0.50:0.95	medium	0.127	0.323
6	0.50:0.95	large	0.189	0.447

TABLE III: Evaluation results of VGG neural network

Evaluation results of VGG16 neural network are visible in table 3. Average precision and average recall has been shown for different iterations with different threshold values being considered.

Iteration	IOU	Area	Average Precision	Average Recall
1	0.50:0.95	all	0.114	0.130
2	0.50	all	0.229	0.253
3	0.75	all	0.103	0.287
4	0.50:0.95	small	0.045	0.168
5	0.50:0.95	medium	0.132	0.317
6	0.50:0.95	large	0.186	0.437

TABLE IV: Evaluation results of resnet50 neural network

Evaluation results of resnet50 neural network are visible in table 4. Average precision and average recall has been shown for different iterations with different threshold values considered.

Iteration	IOU	Area	Average Precision	Average Recall
1	0.50:0.95	all	0.122	0.130
2	0.50	all	0.236	0.260
3	0.75	all	0.118	0.291
4	0.50:0.95	small	0.052	0.136
5	0.50:0.95	medium	0.133	0.327
6	0.50:0.95	large	0.200	0.439

TABLE V: Evaluation results of resnet101 neural network

Evaluation results of resnet101 neural network are visible in table 5. Average precision and average recall has been shown for different iterations with different threshold values considered.

Iteration	IOU	Area	Average Precision	Average Recall
1	0.50:0.95	all	0.118	0.130
2	0.50	all	0.233	0.261
3	0.75	all	0.110	0.290
4	0.50:0.95	small	0.055	0.154
5	0.50:0.95	medium	0.140	0.325
6	0.50:0.95	large	0.186	0.448

TABLE VI: Evaluation results of resnet152 neural network

Evaluation results of resnet152 neural network are visible in table 6. Average precision and average recall has been shown for different iterations with different threshold values considered.

The final result for each image in different iterations that is the relative bounding box pixel values and the corresponding score is visible in the table 7. The first column represents image id, the second column shows the category id or the class id to which the object that was intend to be found in

category id	bbox/0	bbox/1	bbox/2	bbox/3	score
1	504.98	1.53	34.83	40.34	0.6319
1	415.64	52.92	19.09	55.49	0.5869
1	374.25	49.40	11.79	34.78	0.4519
1	372.02	205.18	19.35	81.93	0.3723
1	383.94	45.79	11.08	37.44	0.3461
1	54.86	9.28	75.46	124.53	0.3352
1	384.03	208.72	14.16	70.99	0.3086
1	219.91	35.84	27.39	49.09	0.2910
1	224.65	232.48	23.87	39.12	0.2586
1	191.50	134.42	16.07	53.16	0.22329

TABLE VII: Snapshot of bounding box pixel coordinates for image on different iterations

this target image belongs to. The columns after that bbox0, bbox1, bbox2, bbox3 represents the four relative bounding box pixel values which the model has predicted that means these values represents the four corners of the bounding box where the target object can be found. The final column represents the score related to the bounding box formed with respect to the target object.

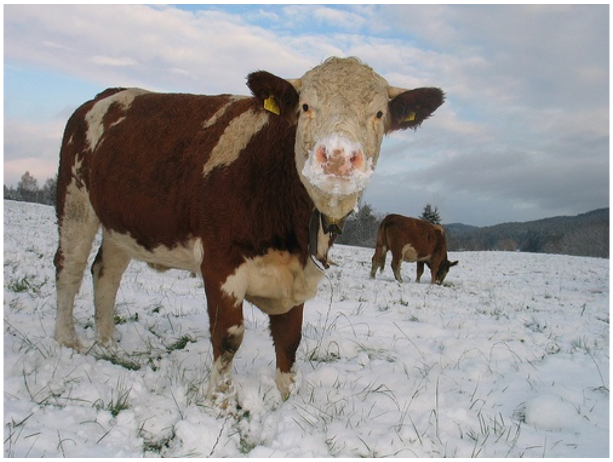

Fig. 3: Before detection of object

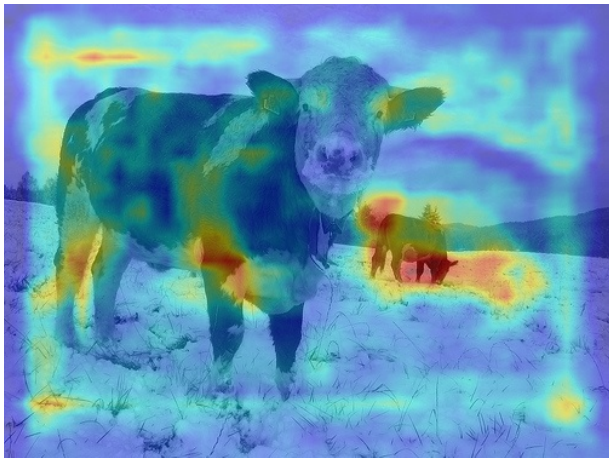

Fig. 4: After detection of object

The result of applying the detection approach can be seen in figure 3 and 4. The bounding box pixel coordinates that were generated using the model localizes the object, in this case: cow.

Performance Metrics	vgg16	resnet50	resnet101	resnet152
Average Precision	0.229	0.231	0.236	0.233
Average Recall	0.437	0.447	0.448	0.439
mAP	0.512	0.561	0.570	0.542

TABLE VIII: Comparison of scores obtained by different networks

The comparison between different neural networks used on the basis of average precision, average recall and mean average precision can be seen in the table 8. It can be seen that resnet101 performed better than all other models followed by resnet50, resnet152 and vgg16.

V. CONCLUSION

In this paper both the proposed co-attention and co-excitation techniques were explored and the correlated evidence revealed by the query-target pairs which makes it generic [20, 21] and not heavily biased to the training data as it not completely depend on training labels.As a result, the proposed method can yield non-local object proposals and uses the co-excitation operation to emphasize important features shared by both the query and the target images. On experimenting with different neural network we found that ResNet101 perform best in comparison to other neural network.

REFERENCES

[1] Leveraging Bottom-Up and Top-Down Attention for Few-Shot Object Detection ,Xianyu Chen, Ming Jiang, and Qi Zhao.

[2] Few-Shot Object Detection with Attention-RPN and Multi-Relation Detector ,Qi Fan,Wei Zhuo,Yu-Wing Tai.

[3] Comparison Network for One-Shot Conditional Object Detection , Tengfei Zhang,, Yue Zhang, Xian Sun, Hao Sun,Menglong Yan, Xue Yang, Kun Fu

[4] Siamese Neural Networks for One-shot Image Recognition , Gregory Koch, Richard Zemel, Ruslan Salakhutdinov

[5] Transfer Learning by Borrowing Examples for Multi-class Object Detection , Joseph J. Lim ,Ruslan Salakhutdinov.

[6] Beyond anchor-based object detector ,Tao Kong, Fuchun Sun, Huaping Liu, Yuning Jiang, and Jianbo Shi.

[7] Detecting objects as paired keypoints ,Hei Law and Jia Deng. Cornernet .

[8] Focal loss for dense object detection,Tsung-Yi Lin, Priya Goyal, Ross B. Girshick, Kaiming He, and Piotr Dollár.

[9] Single shot multibox detector , Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott E. Reed, Cheng-Yang Fu, and Alexander C. Berg.

[10] You only look once: Unified, real-time object detection,Joseph Redmon, Santosh Kumar Divvala, Ross B. Girshick, and Ali Farhadi.

[11] Overfeat: Integrated recognition, localization and detection using convolutional networks,Pierre Sermanet, David Eigen, Xiang Zhang, Michaël Mathieu, Rob Fergus, and Yann LeCun.

[12] Imagenet classification with deep convolutional neural networks by Krizhevsky, Alex, Sutskever, Ilya, and Hinton, Geoffrey E

[13] One shot learning of simple visual concepts by Lake, Brenden M, Salakhutdinov, Ruslan, Gross, Jason, and Tenenbaum, Joshua B

[14] One-shot learning by inverting a compositional causal process by Lake, Brenden M, Salakhutdinov, Ruslan R, and Tenenbaum, Josh

[15] One-shot learning of generative speech concepts Lake, Brenden M, Lee, Chia-ying, Glass, James R, and Tenenbaum, Joshua B.

[16] Zero-shot learning with semantic output codes , Palatucci, Mark, Pomerleau, Dean, Hinton, Geoffrey E, and Mitchell,

[17] Imagenet classification with deep convolutional neural networks by Krizhevsky, Alex, Sutskever, Ilya, and Hinton, Geoffrey E

[18] One shot learning of simple visual concepts , Lake, Brenden M, Salakhutdinov, Ruslan, Gross, Jason, and Tenenbaum, Joshua B

[19] Very deep convolutional networks for large-scale image recognition , Simonyan, Karen and Zisserman, Andrew

[20] One shot learning gesture recognition from rgbd images , Wu, Di, Zhu, Fan, and Shao, Ling

[21] arXiv:1911.12529 [cs.CV]