Project Report : Pattern Recognition(EEL 6825). Object Detection (Using Weakly Supervised learning)

Vikas Chaubey
UFID : 3511 5826
vikas.chaubey@ufl.edu
Department of Computer and Information Science and Engineering
University of Florida
Gainesville, USA

***Abstract*—The project aims at implementing a machine learning solution which can perform task of classification and object detection within a given digital image.Classification and object detection techniques are used to identify objects in a given image, classify them (identify them) and locate them with a bounding box within image. The project focuses on implementing a convolution neural network for classification and uses weakly supervised learning to perform object detection in images using gradient-weighted class activation mapping.**

***Index Terms*—pattern recognition, object detection , classification , localization**

I. Introduction

Object detection is a task in computer vision performed by computer program that involves identifying the presence of any specific category of object, the spatial location of the object within the image, and identifying the category or type of one or more objects in a given image.It is widely used in computer vision tasks such as image annotation, activity recognition, face detection, face recognition, video object co-segmentation etc. When humans look at images or video, we can recognize and locate objects of interest within a matter of moments. The goal of object detection is to replicate this intelligence using a computer program.

A. An Object detection algorithm or computer program has to perform following tasks:

1) Object recognition : Object recognition is a computer vision technique which makes a computer program able to recognize the presence of any specific type of object in the given visual input such as image and video.

2) Object classification : Object classification is a computer vision technique with which a computer program can identify and distinguish multiple categories of objects within the visual input data such as images and video.These different categories of objects are called classes.An object classification algorithm can differentiate among different classes of objects present in the input data such as images and videos.

3) Localization : Localization means spotting the objects locally in the input data i.e. It is a computer vision technique which allows computer program or an algorithm to locate the objects precisely in the given input data such as images and videos.As part of localization process the computer program is supposed to identify the spatial location of the object in the given input data and mark the objects location using a bounding box around the objects position in the given image or video.

B. Different terminologies used in the Object detection:

1) Class : The categories of different types of objects that has to be identified by the object detection algorithm are called classes. for example in this project the Algorithm is trained to identify five different types of animal categories which are dog, cat, mouse, bird and frog.Hence the object detection algorithm works can identify and classify five different classes of animals.

2) Label : The name of the category or a class is called a label for example The object detection algorithm in this project can classify five different classes of animal categories.The name of these animal categories are called labels which are dog, cat, bird , mouse and frog.The Classification output of the object detection algorithm is in the form of labels.

3) Data Annotation : In order to train the object detection algorithm for localization , the training data has to be prepared in a specific manner which involves data annotation i.e. the in the training data different objects(labels or classes) are annotated (their position is defined with bounding boxes in given images or input data).Then this data is used to train the algorithm for localization.This process is called data annotation.

C. Different types of Object detection:

Multi Class Single label : In this category of Object detection the algorithm identifies only one class among multiple classes i.e. for the given image the algorithm identifies only one single label.

Multi Class Multi label : In this category of Object detection the algorithm can identifies multiple classes among in the given input data i.e. for the given image the algorithm identifies multiple objects or labels.

This project uses deep learning convolution neural network to perform image classification(multi class single label classification) and localize the detected object within a bounding box in the image using concepts of weakly supervised learning and gradient-weighted class activation mapping.The project uses convolution neural network for implementing the solution because the multiple inbuilt neural network layers are very good at learning image features , In order to implement a weakly supervised learning solution for object localization which is the next part of the problem, gradient-weighted class activation mapping makes use of the features learned by the last neural network layers.hence a deep learning solution was preferred over traditional machine learning approaches.

A general example of object detection :

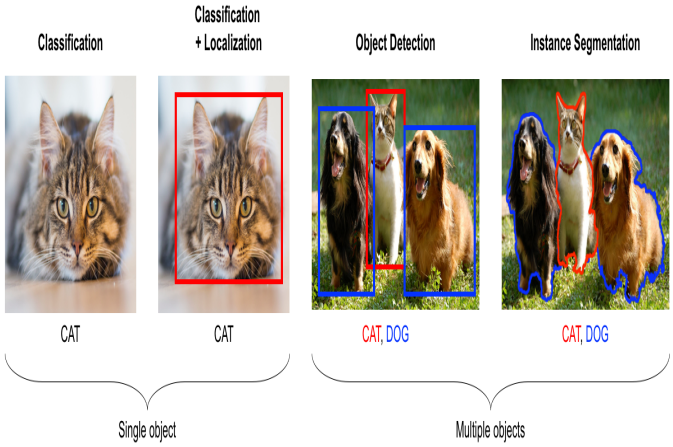

Fig. 1. Classification, Localization ,Object detection,Instance Segmentation Within a image.

II. Motivation

Computer vision and image recognition are integral parts of artificial intelligence.The advancements in this field of AI has has created wide range of application in real world which is helping to reshape the future of human living forever for example the biggest displays of computer vision are coming from the automotive industry in form of self driving cars where object detection is used extensively, another example would be application of computer vision techniques in augmented reality in gaming and mobile phone applications, we have seen that how augmented reality can interestingly en chance user experience several folds with games like "Pokemon go". There is a wide range of applications which are very futuristic and fascinating such as Delivery drones (being developed by Amazon), Pedestrian tracking and face recognition. Not just that computer vision has wide applications in the field of healthcare for example computer vision algorithms can help automate tasks such as detecting cancerous moles in skin images or finding symptoms in x-ray and MRI scans.Hence the main reason and motivation behind choosing object detection for my final project is that applications of computer vision in different areas have always fascinated me.Also doing that project allowed me to use the concepts that i learned in the class such as image classification to identify different classes in given input data, other very important concept which proved really helpful was techniques used for optimization which has proved to be very useful to minimize loss while training the project algorithm other important concepts such as feature engineering , model training, testing and model over fitting have also helped me immensely while developing the project hence it allowed me to implement those concepts practically which has helped me to understand those concepts better.

III. Description

A. Problem Statement and Formulation of Problem :

The project aims at implementing a machine learning solution which can perform task of classification and localization within a given digital image.Classification and localization techniques are used to identify objects in a given image, classify them (identify them) and locate them with a bounding box within image.

Classification is the process of predicting the class of given data points. Classes are sometimes called as targets/ labels or categories. Classification predictive modeling is the task of approximating a mapping function (f) from input variables (X) to discrete output variables (y).Classification belongs to the category of supervised learning where the targets also provided with the input data.

Object localization is a process which predicts the object in an image as well as its boundaries.object localization aims to locate the main (or most visible) object in an image .

In this project Classification part of the problem is resolved by a special type of deep learning neural network called convolution neural network, ideally for object detection and localization neural networks are trained for both tasks (classification and localization), but training such image classifiers capable of doing object detection and localization both is a very costly process since it requires labelling(class labels) and annotating data(bounding boxes around objects) which could be very expensive if the data-sets are very large.

Hence in this project the neural network is trained using only labelled data and in order to perform localization of objects weakly supervised learning is used i.e. while training the neural network data will not be annotated (drawing bounding boxes around objects) , but using outcomes obtained from the neural network convolution layers , gradient-weighted class activation maps will be generated in order to obtain localization of detected class and with the help of these gradient class activation maps bounding boxes could be drawn around the detected object in the input image.

Since class activation maps make use of outcomes obtained from the last convolution layers of neural network ,In order to Localize objects in image correctly it is necessary that neural network class prediction is accurate.Hence the project problem is mainly to develop a highly accurate neural network for image classification. if the class prediction is accurate the class activation map generated by GRAD-CAM algorithm will also be accurate and hence drawn bound boxes will be at precise locations where the objects are present.

The resultant classifier is trained to classify five different classes of animals , the five different animal classes are : Dog, Bird, Frog, Cat , Mouse. the classifier would be a multi class single label classifier i.e. the classifier can predict up to five classes but it predicts one class label per image input.It can not predict multiple class labels within same image.

A convolution neural network is consisted of many layers the initial layers are called convolution layers which perform function of learning image features like edge detection and shapes etc. These layers perform convolution of kernels or filters(f) with the Image or vidual data(V) and the outputs are passed to next consecutive layer.the convolution output could be defined as :

$$C[p,q] = (V * f)[p,q] = \sum_a \sum_b V[p,q] f[p-a, q-b] \quad (1)$$

In convolution neural network the optimized parameters or weights are obtained by the process of forward and backward propagation.Forward propagation works similar as it works in the feed forward networks.During forward propagation The output of the previous layer(Y[k-1]) is convoluted with the weight parameters(P)(filters) of the current layer and this value is summed with bias(b) the overall summation forms the intermediate output value for the current layer then this outcome is fed to an non linear activation function.The output of the activation function is the final output of the convolution layer. here K represent the layers of convolution neural network.

$$T^k = P^k * Y^{k-1} + b^k \quad (2)$$

$$Y^k = a^k(T^k) \quad (3)$$

The output obtained fro the forward propagation is then fed to a loss function, during back propagation loss function derivatives are derived in order to minimize the loss function.during back propagation the process starts from the last layer and continues till the very first layer while obtaining obtaining parameter values for each layer of the convolution neural network.

$$dY^k = \frac{\partial Loss}{\partial Y^k} \quad (4)$$

$$dT^k = \frac{\partial Loss}{\partial T^k} \quad (5)$$

$$dP^k = \frac{\partial Loss}{\partial P^k} \quad (6)$$

$$db^k = \frac{\partial Loss}{\partial b^k} \quad (7)$$

$$dT^k = dY^k * a'(T^k) \quad (8)$$

The loss is a log loss function which is determined as, this log loss function has to me minimized in order to make the prediction model more accurate :

$$Loss(y, \hat{y}) = \sum_{j=0}^{a} \sum_{i=0}^{b} (y_{pq} log(\hat{y}_{pq})) \quad (9)$$

where y represents the actual outcome and and y hat represents the prediction made by the classifier.

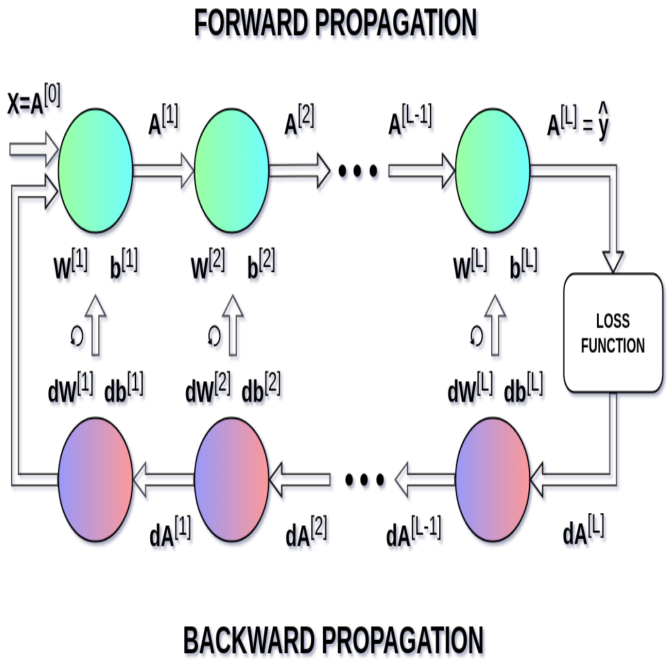

Fig. 2. Forward and Backward Pass in a convolution neural network

B. Algorithms

The research in the field of computer vision techniques has proposed various ways and algorithms to solve the problem of object detection.These different algorithm solutions could be mainly divided into two categories: 1) Traditional machine Learning based computer vision approaches : The traditional machine learning approaches involve computer vision techniques such as as feature descriptors SIFT(Scale-invariant feature transform) ,SURF (Speed-ed-Up Robust Features),BRIEF etc. for object detection.All these techniques require feature extraction from the data which are also called descriptors or informative segments in a image which define its context for the task of image classification.Several In order to learn descriptors of the image in traditional computer vision techniques the CV algorithms such as edge detection, corner detection or threshold segmentation are used.Initially for a given category of input data or class as many important descriptors are evaluated in order to be able to precisely define the class for a computer program.These features respected to each class are fed into the object detection algorithm based on those feature based definition the object detection algorithm makes prediction for the given input data.The downside of using these algorithms are that they need feature engineering in order to extract descriptors from images, These algorithms are not capable of learning those features by themselves, since human intelligence is required to implement feature engineering these

algorithms are very prone to errors because humans can make errors while extracting descriptors from the available data.Also preparation of descriptors consumes a lot of time.

The second category of the object detection algorithms fall into deep learning based approaches.Deep learning is the branch of machine learning which uses neural networks in order to solve the standard machine learning problems.These neural networks function in the same manner a human brain does.Neural networks are designed in similar way as a human brain in order to replicate the learning and functionality of human brain.Vision is easily understood natural phenomenon for humans.Humans can see objects through eyes and human brain can easily comprehend this visual data in order to make sense of what the eyes are seeing.hence neural networks try to replicate the human vision process in order to implement the effective computer vision solutions. some effective neural networks which are used for object detection are CNN(convolution neural networks), R-CNN, Fast R-CNN, Faster R-CNN, Retina-Net, Deform able convolutional networks etc.Some deep learning based approaches such as YOLO(V2), YOLO(V3) are some of the most popular deep learning solutions for object detection.The advantage of using deep learning solution for computer vision problems is that they don't require feature engineering, The neural networks can learn the image features on their own from the input data, neural networks are capable of learning the features and make sense of information by themselves just like human brains.Hence neural networks have proved to be very accurate while doing object detection.

Among all deep learning based neural networks convolution neural networks are the most popular to solve computer vision problems because these special types of neural networks perform very well in object detection and image classification tasks. This is because they are designed after visual cortex of the human brain which enables humans to see and comprehend the visual information seen by eyes.Although the results obtained by such neural networks are very promising but the main challenge lies in the preparation of the data which is used to train these neural networks.As discussed earlier neural networks are capable of extracting feature information , pattern for different classes from the given data, but in order to train them properly large amount of data sets are required.Especially in case of object detection data preparation task becomes more cumbersome because the data has to be labelled as well as annotated with bounding boxes in order to teach the neural network classification(information using labels) and localization(information using annotated bounding boxes).also annotations produced by humans are prone to errors.Since convectional neural networks require thorough data preparation for training for the task of object detection , the data preparation could prove to be expensive process if the size of the data sets are really large.

Hence In order to reduce this training effort to prepare data with labels and annotated bounding boxes,this project focuses on a special type of training technique which is called "weakly supervised learning".In these techniques labelled data without bounding box annotation could be used for training of models which can reduce data preparation effort and reduce development costs. this project implements Object detection using weakly supervised learning techniques hence training data sets are not prepared with annotations(bounding boxes).The object detection problem has mainly two components to it the first is classification and second is object localization in the image.In this project the classification part of the problem is handled by "convolution neural network" and localization is done using "gradient-weighted class activation mapping" algorithm (also called Grad-Cam Algorithm) which enables this algorithm to be weakly supervised for localization task.

1) Convolution Neural Networks (CNN): A convolution neural network (CNN) is a type of deep neural network, It is most commonly used to do visual imagery such as images and videos. As the name suggests convolution neural networks use a convolution in place of matrix multiplication in at least one of their layer. Convolution is a specialized kind of linear operation..convolution neural networks have wide range of applications in image and video recognition, recommend-er systems, image classification, medical image analysis, natural language processing and financial time series etc.

Why traditional Neural networks are not good for image processing? Traditional neural network such as multi layer perceptron are densely connected neural networks, these densly connected neural networks are structured in such a way that each neuron unit is connected to every other neuron unit of the network.A neuron in the feed forward network is a mathematical representation of function of actual human brain neurons.These type of neural networks are modeled according to the structure of human brain.since these networks are fully connected hence these networks are prone to getting over-fit while training and they do not generalize well on the Image data.These types of networks are good and work well on a limited set of defined features for solving problems such as classification problem. But in case of images and visual data this approach is very expensive.In case of large input data such as images the feed forward networks will have to contain a lot of neuron units hence the the parameters to determine in order to produce a balanced model become very difficult because of the large number of parameters.Hence training such networks for large input data becomes very difficult.

Why convolution neural networks work very well with images and visual data? The process of Image Recognition is the process of identification of image by a computer or machine that retrieves information and context from a given image to understand what image depicts.Human brains are very good at interpreting the visuals easily. Human brains can inherently generate understanding of any visual data such as a view, image or video. In most cases, Human brains do not struggle or need to do a conscious study of the visual object to make sense of it.Interpretation of visuals take only fraction of seconds in human brain , While human brain is comprehending and making sense of visuals seen by eyes a complex cognitive process occurs in the visual cortex of the brain.

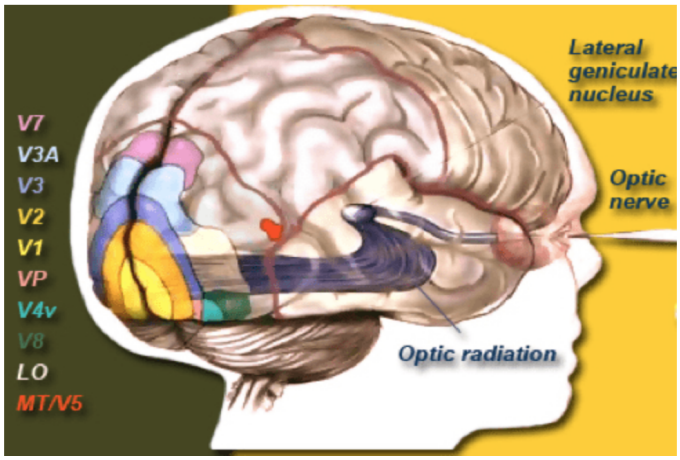

Fig. 3. Representation of Human Visual Cortex.

Figure 1 presents a visual presentation of of the visual cortex, visual cortex is divided into multiple layers (to make it understandable lets assume these layers are V1-V8 which are strongly connected), and these layers are responsible to process visual information coming from the eyes. When human eyes sees a perceivable object then the image is projected on retina, in turn optic nerves take those signals to the visual cortex of the brain which is formed by multiple layers, the initial layers of the visual cortex are responsible to understand the high level features of the visual being seen by eye such as edges, shapes etc. on the other hand deep layers of the visual cortex comprehend the context and more detailed information.

Convolution neural networks are designed as per the functioning of visual cortex of the human brain just like the visual cortex convolution neural networks are also comprised of multiple layers.Many studies performed on the visual cortex suggest that when eye sees any visual then specific neurons in the specific area of brain are fired.it was seen that not all neurons are activated at once and visual cortex function in a layered manner where the different layers of the cortex are activated sequentially, convolution neural network make use of this information.hence in convolution neural network multiple layers of the network work sequentially and all neural units are not interconnected , a single neuron unit connects to certain other neural units.this kind of architecture imitates the visual cortex in order to do effective computation with visual data with less effort.Since the convolution neural networks imitates the visual cortex of human brain hence they function very well with visual data.

Architecture of convolution neural network mainly comprised of following layers: 1) Convolution Layer: As the name suggests these layers perform the operation of convolution between input image and various filters which are also called kernels.This convolution operation generates various class activation maps for different kernels which enable these layers to learn the image features.

2) Pooling Layers : The class activation maps generated in the convolution layer are fed to the pooling layers.this layers perform the function of max pooling, down sampling and sub sampling.

3) Dense layers : These layers are the last layer in the CNN architecture, results obtained in the previous layers are finally fed to the dense layers, these layers are responsible to process the outcomes and do the final classification using activation functions.

Block Diagram of different layers of a convolution Neural Network :

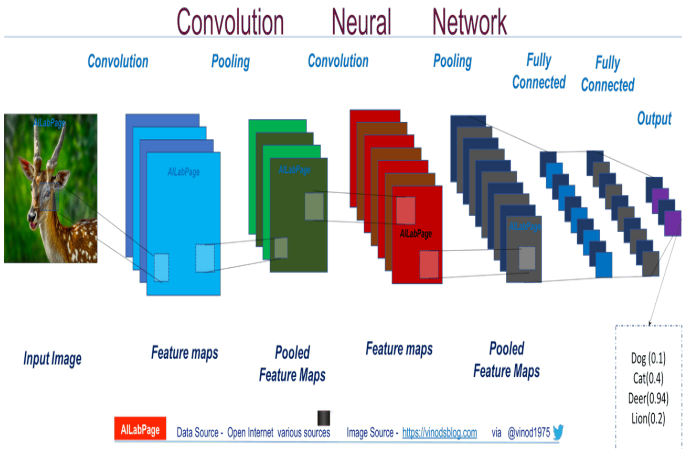

Fig. 4. Block Diagram : Different layers in Convolution Neural Network.

2) Gradient-Weighted Class Activation Mapping(Grad-Cam Algorithm): The second part of the project problem statement is object localization within the image, when the image is classified by the convolutional neural network the second task is to locate the detected object within the image.generally for object detection and localization neural networks are trained for both tasks (classification and localization), but training such image classifiers capable of doing object detection and localization both is a very costly process since it requires labelling(class labels) and annotating data(bounding boxes around objects) which could be very expensive if the datasets are very large.

In this project the localization is implemented using class activation map (CAM) visualization techniques.Class activation maps is a simple technique used by CNNs in order to identify different edge regions in a given image using those regions CNN can determine the class of the object in the given image.In this process the algorithm which is responsible for localization make use of the feature maps produced by the last convolution layer of the neural network, these feature maps are used to generate the heat maps of the given input image.These heat maps are generated by checking what regions of image were activated and how similar they are with respect to the considered class.The localization based on class activation mapping could be implemented using “Grad-CAM(Gradient-Weighted Class activation mapping) algorithm which can draw Visual explanations from deep neural networks via Gradient-based Localization.Using this algorithm localization could be done without training the model with annotated training

data with bounding boxes to implement model with "weakly supervised learning"

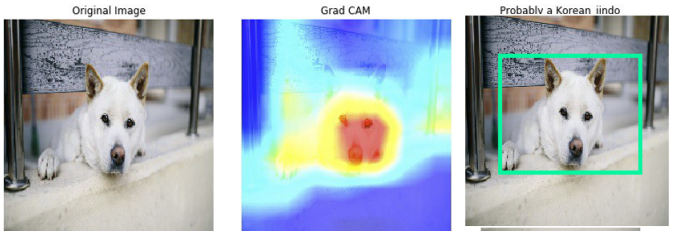

Fig. 5. Gradient-weighted Class Activation Maps - Used for Localization

C. Implementation

1) Programming Language and System Specs: The project is implemented using following components:

1)programming Language : Python 3.7
2)Keras : A python Library to implement neural network and training model
3)Matplotlib : A python library to visualize data using plots
4)PyQt5 : A python library to make desktop based Graphical user interfaces
5)Numpy : A python library to perform mathematical operations
6)Open CV2 : a python library to process image data
7)Operating System : MacOs Mojave
8)Training platform : GPU (Radeon Pro 560X 4 GB)

2) Data Set Preparation: The data sets is prepared using the images from imagenet. The imagenet project is a large visual database designed for use in visual object recognition software research consisting more than 14 million images across 20000 classes.Imagenet provides URls to download images for free using class wordnetId. Imagenet provides good quality pictures which are sufficient to implement the project.Hence the project was implemented using data set prepared by Imagenet images.This new data set consists of five different classes of animals which are Dog,Cat,Mouse,bird and Frog. Total 7500 images were downloaded across all 5 classes , 1500 images for each class(Dog,Cat,Mouse,bird and Frog)and this data set was finally used to train the project model.

3) Implementation of convolution neural network(CNN) architecture: The project uses the VGG-16 architecture which is a type of convolution neural network and is very popular for tasks related to computer vision because of its high accuracy. VGG-16 architecture contains 13 convolution layers which perform convolution operation , apart from that there are 5 Max Pooling layers which perform max pooling and operations like down sampling for the outcomes obtained by convolution layers and 3 Dense layers which are responsible to generate final classification output.Hence the convolution neural networks consist of total 21 layers out of which only 16 layers are weighted.

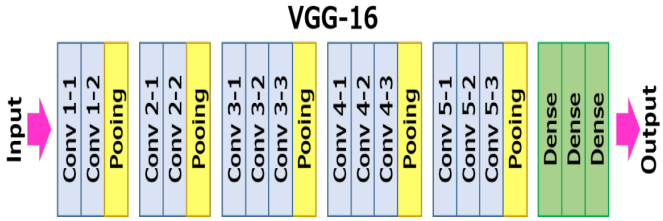

Fig. 6. Block diagram : Architecture of VGG16 Convolution Neural Network.

4) Implementation of Grad-Cam algorithm: Grad-Cam algorithm for localization of detected objects is implemented using Open CV which is a python library to process images, The the steps involved in implementation of Grad cam are:

1) Obtain the output of the last convolution neural network for the given input image
2) Obtain the classifier output from CNN to identify the output class for the input image
3) generate the heatmap for the input image by using above two parameters
4) Use the heatmap to draw bounding box around the activated region (which is the object in the image)

5) Training: For training of the model the data set is divided into 80-20 split, i.e. 80 percent of the images out of all available images in the data set are used for training remaining 20 percent images are used for validation testing of the model. Out of total 7500 images for all 5 classes (Dog,Cat,Bird,Frog and mouse), 6500 images were used for training remaining 1000 images were used for validation.the training data set is comprised of 1300 images per class (totalling 6500 images).While the validation data set consists of 200 images per class (totalling 1000 images). Training was done for 50 epocs which took around 53 hours on local machine GPU.

IV. Evaluation

Evaluation of the model is based on the training and testing accuracy statistics obtained while training the model on training data sets and performing validation on the validation data sets.The accuracy stat is obtained for bothe the set when the model is trained.One more parameter to evaluate the accuracy of the model is loss function for training and validation data sets.In order to perform the accuracy check for localization we can only rely on manual testing , in order to obtain localization stat , the annotated data set is required with which the results can be compared since its not available manual testing is done for localization.

A. Training and Validation Accuracy

The accuracy obtained for model on the training data set is 79 percent and The accuracy obtained for model on the validation data set is 53 percent for total 50 epochs.The accuracy plot shows the improvement in training and testing validation set as the training continued for 50 epochs.

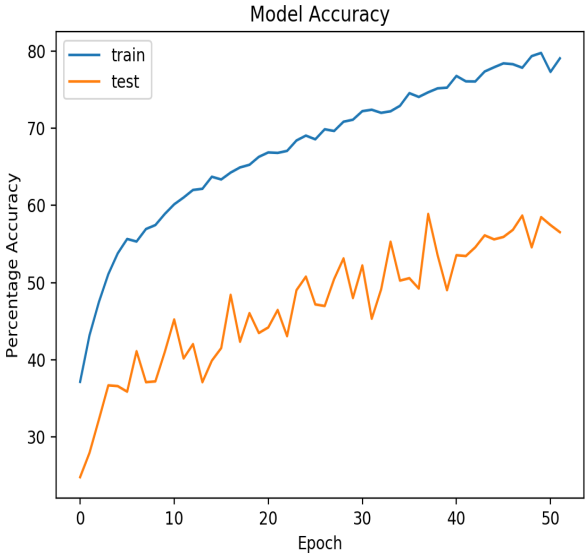

Fig. 7. Training and Validation Accuracy Plot

2) The loss obtained for the model on the training data set is nearly : 0.55 and The loss obtained for the model on the validation data set is nearly : 1.55 for total 50 epochs. The Loss plot shows the improvement in training and testing validation data set loss values as the training continued for 50 epochs.

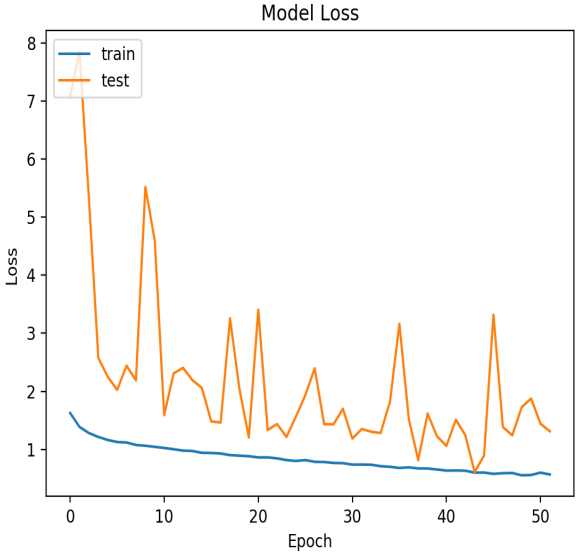

Fig. 8. Training and Validation Loss Plot

B. Results Obtained on Test data

Correct classification and localization results on test data :

Dog Class in the image classified correctly and Localization is done with correct bounding box location :

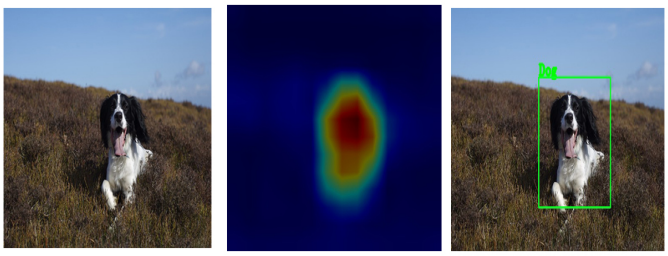

Fig. 9. Correct Classification and Localization for class dog

Cat Class in the image classified correctly and Localization is done with correct bounding box location :

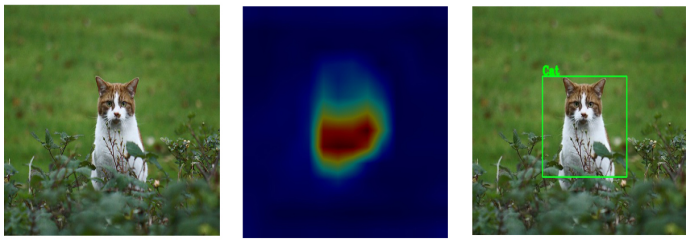

Fig. 10. Correct Classification and Localization for class cat

Bird Class in the image classified correctly and Localization is done with correct bounding box location :

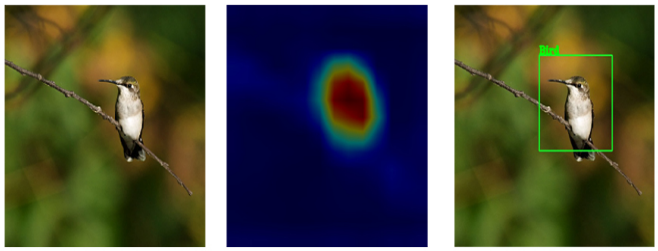

Fig. 11. Correct Classification and Localization for class bird

2) Correct classification But Incorrect localization results on test data :

Bird Class in the image classified correctly but Localization is incorrect :

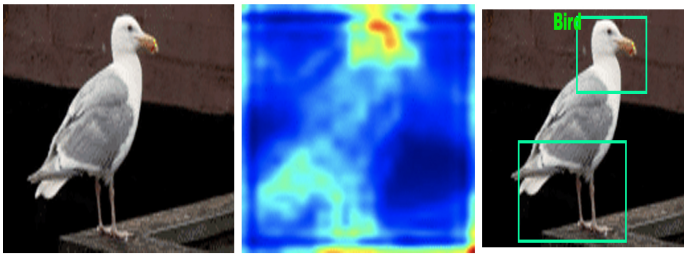

Fig. 12. Correct Classification , Incorrect Localization for class bird

Dog Class in the image classified correctly but Localization is incorrect :

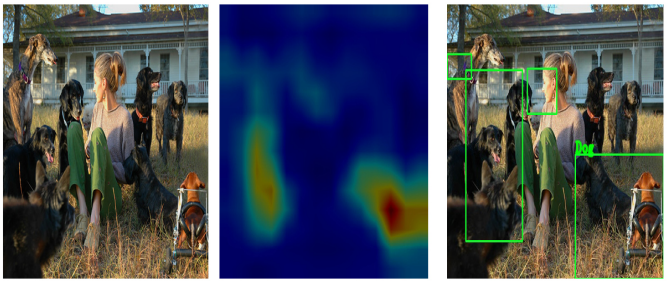

Fig. 13. Correct Classification, Incorrect Localization for class dog

Cat Class in the image classified correctly but Localization is incorrect :

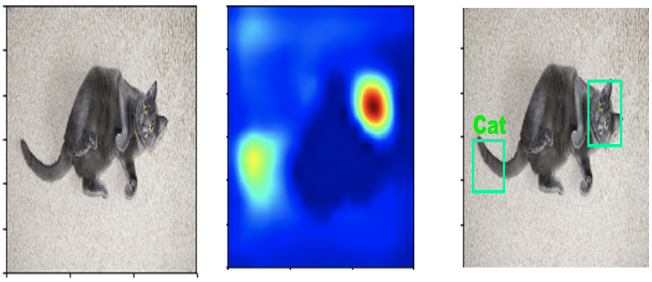

Fig. 14. Correct Classification , Incorrect Localization for class cat

3) Incorrect classification and Incorrect localization results on test data : All the test images(around 50 images) which were tested manually had correct classification results , no images were found with output with incorrect classification results hence this case could not be seen in manual testing with images, however incorrect localization results were seen in case2. The classifier accuracy is very good in predicting the image class but it performance is average when it comes to localization.

V. Related work

Some of the traditional machine learning approaches involve computer vision techniques such as as feature descriptors SIFT(Scale-invariant feature transform) ,SURF (Speed-ed-Up Robust Features),BRIEF etc. for object detection.All these techniques require feature extraction from the data which are also called descriptors or informative segments in a image which define its context for the task of image classification.In order to learn descriptors of the image in traditional computer vision techniques the CV algorithms such as edge detection, corner detection or threshold segmentation are used. Pros and Cons of traditional Machine learning based object detection algorithms: Pros: 1) Since feature engineering or descriptors are extracted by human intelligence which are fed to algorithm later , hence the training time of the algorithms reduce significantly 2) with good feature engineering algorithms can perform well with good accuracy 3) Algorithms do not require very large data sets in order to learn patterns and features, hence when data sets are very small these algorithms prove to be very useful. 4) These algorithms do not require very high hardware processing power such as GPU units in order to learn the patterns. Cons: 1) Feature extraction is not inbuilt and is done by human intelligence first which is later fed to the algorithm 2) The accuracy of the algorithm highly depends upon the correctness of feature extraction as it is done manually it is prone to human made errors.

The second category of the object detection algorithms fall into deep learning based approaches.Deep learning is the branch of machine learning which uses neural networks in order to solve the standard machine learning problems.some effective neural networks which are used for object detection are CNN(convolution neural networks), R-CNN, Fast R- CNN, Faster R-CNN, Retina-Net, Deform able convolution networks etc.Some deep learning based approaches such as YOLO(V2), YOLO(V3) are some of the most popular deep learning solutions for object detection.Pros and Cons of traditional deep learning based object detection algorithms: Pros : 1) These algorithms are very fast in producing results from input data 2) Feature extraction is handled by algorithms only as neural networks are self sufficient to learn patterns and features from raw data by themselves. 3)These algorithms are very accurate in terms of object classification and localization hence they have very good accuracy Cons: 1) The deep learning based neural networks require very large data sets for training hence in the scenarios where large data sets are not available traditional machine learning approaches are preferred. 2) In order to train deep learning based neural networks for object detection, requires labelling(class labels) and annotating data(bounding boxes around objects) which could be very expensive if the data-sets are very large.labelling and annotating data requires significant effort. 3) Since Deep learning neural networks do feature extraction by themselves hence training such networks require more time for training. 4) For training large amount of processing power is required in order to process large data sets.

The goal of this project was to implement a deep learning based convolution neural network which can perform task of classification and localization by utilizing weakly supervised learning in order to minimize the data preparation effort which is significantly more in deep learning approaches.Deep

learning based neural networks are trained for both tasks classification and localization hence the training data has to be prepared for both tasks hence data labeling is done in order to teach a neural network about different classes and data annotation is done in order to teach a neural network about localization.Data annotation is a heavy data preparation task.In this project only labelled data is used to train classifier for classification and localization is done using class activation maps generated by the convolution layers of neural network using grad- cam algorithm.Pros and Cons of training a deep learning neural network using weakly supervised learning(Grad-Cam Algorithm): Pros: 1) The trained model is fast in producing results 2) Feature extraction is handled by neural network hence accuracy is not dependent upon human effort. 3) data preparation eff fort is less because data annotation is not required for training of classifier only labelled data is sufficient. Cons: 1) In order to obtain high accuracy Large data sets are required 2) The accuracy of localization is highly dependent on the accuracy of classification by the convolution neural network, because localization is done using Grad-Cam algorithm which use class activation maps generated by last convolution layers in the neural network. 3) In order to produce correct localization results classifier should be highly accurate.

VI. Summary and Conclusions

After implementation of the deep learning based convolution neural network using weakly supervised learning in order to solve the problem of object detection (Classification and localization) , following conclusions were made:

1) This approach of training a neural network is good in scenarios where data sets are very very large and its very difficult to annotate data to train model for localization.If data preparation cost is very high then this approach could be used to implement a deep learning based solution because it does not require annotated data and it can still perform object localization along with classification.

2) After implementation of the model it was observed that Trained model is fast in producing results on test data.

3) Implementation of neural networks require large data sets for training and large amount of processing power in order to train the model.Also training is a time consuming process, In this project implementation total 52 hours of training across 7500 images from imagenet data set across five different classes was involved.hence training neural networks take significant effort.

4) The trained model (trained for object detection in multi class single label paradigm using 7500 images across 5 different classes) performs the task of classification with very high accuracy.It was expected because the number of classes are minimal, but The focus of the project was to implement object localization using the classifier output without explicitly training it for localization.

5) the trained model localization accuracy is below average, for images with single class where the image has presence of only one category the localization works fine but when the image is complicated and has multiple objects even if the classification is done accurately localization results are poor.

6) The localization results depend highly upon the classifier results hence in order to obtain good localization results the classifier should be trained with high accuracy.

References

[1] R. Chauhan, K. K. Ghanshala and R. C. Joshi, "Convolutional Neural Network (CNN) for Image Detection and Recognition," 2018 First International Conference on Secure Cyber Computing and Communication (ICSCCC), Jalandhar, India, 2018, pp. 278-282.

[2] R. Girshick, "Fast R-CNN," 2015 IEEE International Conference on Computer Vision (ICCV), Santiago, 2015, pp. 1440-1448.

[3] S. Ren, K. He, R. Girshick and J. Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 39, no. 6, pp. 1137-1149, 1 June 2017.

[4] J. Redmon, S. Divvala, R. Girshick and A. Farhadi, "You Only Look Once: Unified, Real-Time Object Detection," 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, 2016, pp. 779-788.

[5] Francois Chollet. 2017. Deep Learning with Python (1st. ed.). Manning Publications Co., USA.

[6] R. Li and J. Yang, "Improved YOLOv2 Object Detection Model," 2018 6th International Conference on Multimedia Computing and Systems (ICMCS), Rabat, 2018, pp. 1-6.

[7] Y. Lee, C. Lee, H. Lee and J. Kim, "Fast Detection of Objects Using a YOLOv3 Network for a Vending Machine," 2019 IEEE International Conference on Artificial Intelligence Circuits and Systems (AICAS), Hsinchu, Taiwan, 2019, pp. 132-136.

[8] J. Deng, W. Dong, R. Socher, L. Li, Kai Li and Li Fei-Fei, "ImageNet: A large-scale hierarchical image database," 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, 2009, pp. 248-255.

[9] S. A. K. Tareen and Z. Saleem, "A comparative analysis of SIFT, SURF, KAZE, AKAZE, ORB, and BRISK," 2018 International Conference on Computing, Mathematics and Engineering Technologies (iCoMET), Sukkur, 2018, pp. 1-10.

[10] M. Shaha and M. Pawar, "Transfer Learning for Image Classification," 2018 Second International Conference on Electronics, Communication and Aerospace Technology (ICECA), Coimbatore, 2018, pp. 656-660.

[11] H. Qassim, A. Verma and D. Feinzimer, "Compressed residual-VGG16 CNN model for big data places image recognition," 2018 IEEE 8th Annual Computing and Communication Workshop and Conference (CCWC), Las Vegas, NV, 2018, pp. 169-175.

[12] T. Tagaris, M. Sdraka and A. Stafylopatis, "High-Resolution Class Activation Mapping," 2019 IEEE International Conference on Image Processing (ICIP), Taipei, Taiwan, 2019, pp. 4514-4518.

[13] R. Yang, X. Xu, Z. Xu, C. Ding and F. Pu, "A Class Activation Mapping Guided Adversarial Training Method for Land-Use Classification and Object Detection," IGARSS 2019 - 2019 IEEE International Geoscience and Remote Sensing Symposium, Yokohama, Japan, 2019, pp. 9474-9477.

[14] A. Kwaśniewska, J. Rumiński and P. Rad, "Deep features class activation map for thermal face detection and tracking," 2017 10th International Conference on Human System Interactions (HSI), Ulsan, 2017, pp. 41-47.

[15] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh and D. Batra, "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization," 2017 IEEE International Conference on Computer Vision (ICCV), Venice, 2017, pp. 618-626.

[16] A. Bergamo, L. Bazzani, D. Anguelov, and L. Torresani. Self-taught object localization with deep networks. arXiv preprint arXiv:1409.3964, 2014. 1, 2

[17] P . Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun. Overfeat: Integrated recognition, localization and detection using convolutional networks. arXiv preprint arXiv:1312.6229, 2013.

[18] A. Chattopadhay, A. Sarkar, P. Howlader and V. N. Balasubramanian,"Grad-CAM++: Generalized Gradient- Based Visual Explanations for Deep Convolutional Networks," 2018 IEEE Winter Conference on Applications of Computer Vision (WACV), Lake Tahoe, NV , 2018, pp. 839-847.

[19] http://gradcam.cloudcv.org/
[20] keras. https://keras.io/
[21] Open CV. https://opencv.org/
[22] https://www.tensorflow.org/
[23] https://numpy.org/
[24] https://www.python.org/