CASA-Crowd: A Context-Aware Scale Aggregation CNN-Based Crowd Counting Technique

AUTHORS :

NAVEED ILYAS 1

ASHFAQ AHMAD 2 , (Student Member, IEEE),

AND KISEON KIM 1 , (Senior Member, IEEE)

1School of Electrical Engineering and Computer Science, Gwangju Institute of Science and Technology (GIST), Gwangju 61005, South Korea 2School of Electrical Engineering and Computing, The University of Newcastle, Callaghan, NSW 2308, Australia

TEAM MEMBERS :

DEVANSH JAIN (1806174)

ABHISHEK KUMAR (1806107)

YASH VARGANTWAR (1806173)

Published on: 17th December, 2019

INDEX

INTRODUCTION

CASA breaks into context aware and scale aggregation :

Context aware- small big image and irregular image structure In each block, an inception module is included with varying dilation rate to obtain the scale and contextual-aware features. With varying receptive field due to different dilation rate, it is very helpful to cope up the perspective variation issues.

Scale aggregation because of dilution where 3*3 turns to 5*5 or 7*7 and in block 5 -7 concatenation of various dilation rate Conv2D networks is done.

PROBLEM STATEMENT –

CASA-Crowd: A Context-Aware Scale Aggregation CNN-Based Crowd Counting Technique.

OBJECTIVE –

1.) Detection, tracking and counting in low resolution imagesand surveillance videos, where people are represented by only few pixels tall, are issues that yet demand more investigation.

INTRODUCTION

2.) Nonetheless it is very challenging to accurately obtain the count due to severe occlusion, clutter, irregular object distribution and non-uniform object scale , it is possible to solve this problem with the boost of Convolutional Neural Network(CNN).

3.) A Context-aware Scale Aggregation CNN-based Crowd Counting method (CASA-Crowd) to obtain the deep, varying scale and perspective varying features

DATASET DESCRIPTION

SOURCE - https://www.kaggle.com/tthien/shanghaitech-with-people-density-map

DATASET –

THE DATASET CONTAINS TWO PARTS-

- "SHANGHAITECH PART A" contains 482 IMAGES out of which 300 images are used for training the model and the rest 182 images are used to test the model accuracy whereas "SHANGHAITECH PART A" contains 716 IMAGES out of which 400 images are used for training the model and the rest 316 images are used to test the model accuracy.
- WE HAVE USED "SHANGHAITECH PART A" DATASET WHILE TRAINING OUR MODEL.

DATASET CONTAINS–

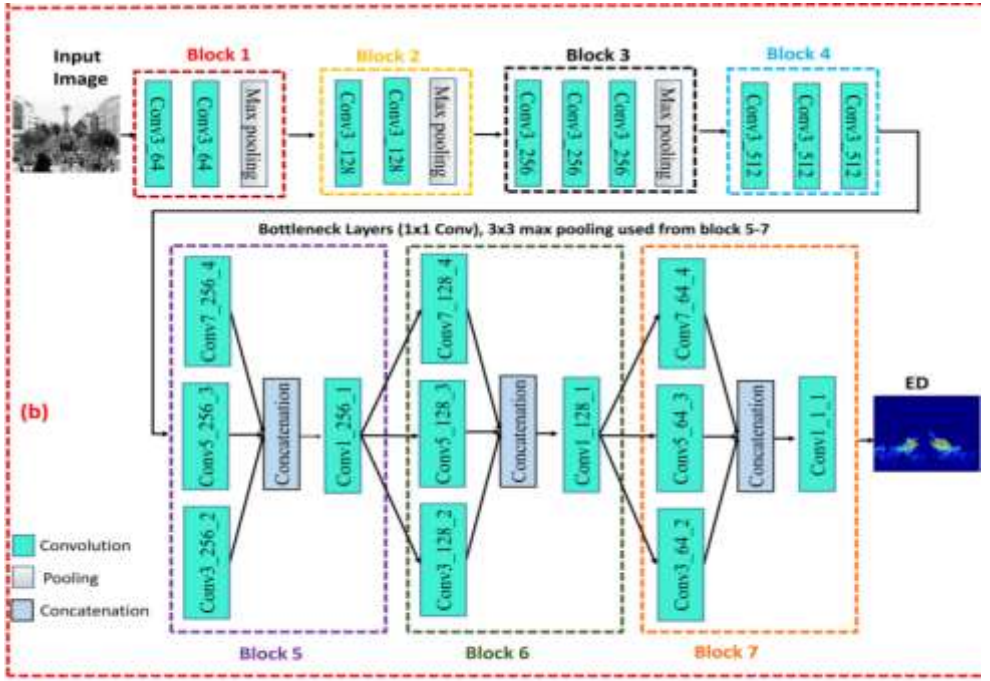
- Images folder which contains the jpg image images , ground-truth which has the .mat format file containing annotated head (coordinate x, y) and ground-truth-h5 which contains people density map of h5 file format.

IMAGE PROCESSING –

- Normalization of image is done where we have created a separate function called "norm_by_imagenet" in which the image if first divided by 255 and then also divided by the mean and standard deviation of each layer of the image.
- Resizing of image is being applied using cv2.resize() where cv2 is the python library and resize is the function of cv2.
- Image is also flipped horizontally and we have created a separate function for that in the utilsimg_proc.py file called "flip_horizontally".

MODEL AND ARCHITECTURE

1) The given picture is of the proposed network in the paper and exactly that is being used in our project with no changes as such. Also a pretrained Model called VGG-16 is used which is also made through a large number of convolutional networking Layers.



IGURE 2. (a) The overview of CASA-Crowd, The deep feature extraction network (DFEN), scale aggregation module with dilated convolutio SAD), (b) The whole architecture consist of two parts: one is deep feature extraction network which consist of 4 blocks, the other is Scale ggregation Module with Dilated Convolution consist of three blocks. The convolutional layers parameters are denoted as "Conv-(kernel ize)-(number of filters)-(dilation rate)", max-pooling layers are conducted over a 2 × 2 pixel window with stride of 2.

MODEL AND ARCHITECTURE

2) The proposed network employs two types of network: a network with smaller and same size of filters (block 1- block 4) inspired from VGG-16 named as DFEN. This network is capable of extracting the simple to complex deeper features. When an input is fed to the architecture, it passed through block 1 to block 4 successively. Here, block 5-7 are named as SAD. In each block, an inception module is included with varying dilation rate to obtain the scale and contextual-aware features. With varying receptive field due to different dilation rate, it is very helpful to cope up the perspective variation issues.

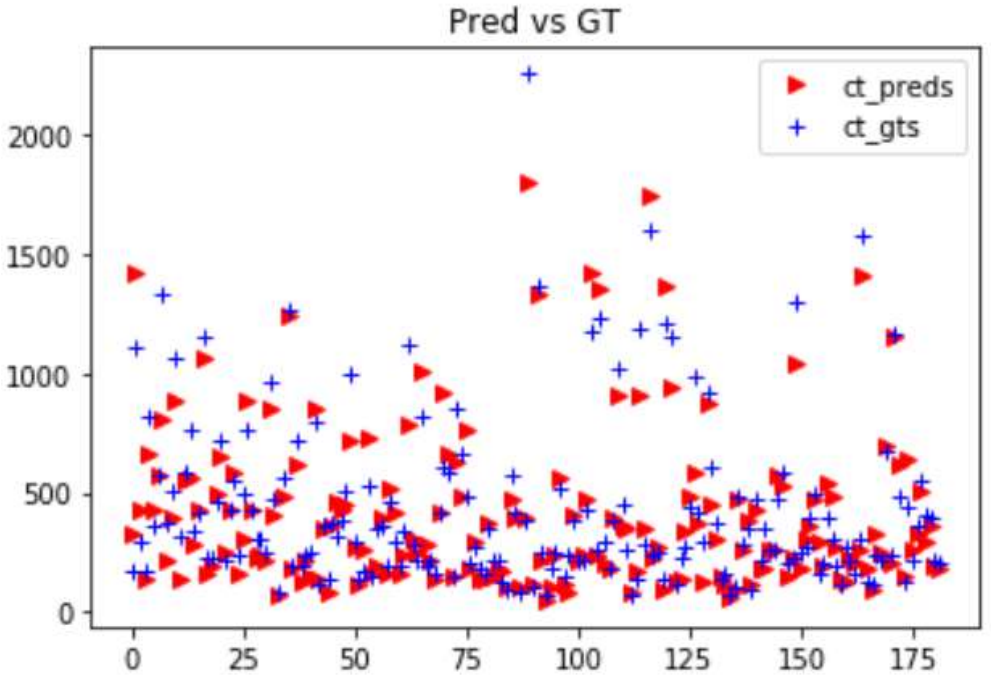
3) Due to fewer number of images, it has been conducted by many deep learning models, to use pretrained models to avoid overfitting. We choose VGG-16 as the front end of CASA-Crowd due to its strong transfer learning ability. In this way, it is a flexible architecture to concatenate with SAD for density estimation.

RESULT OBTAINED

IN THE GIVEN IMAGE WE CAN SEE THE GROUND TRUTH AND OUR PREDICTED VALUES OF CROWD COUNTING.

CODE:

```
plt.plot(ct_preds, 'r>')

plt.plot(ct_gts, 'b+')

plt.legend(['ct_preds', 'ct_gts'])

plt.title('Pred vs GT')

plt.show()
```

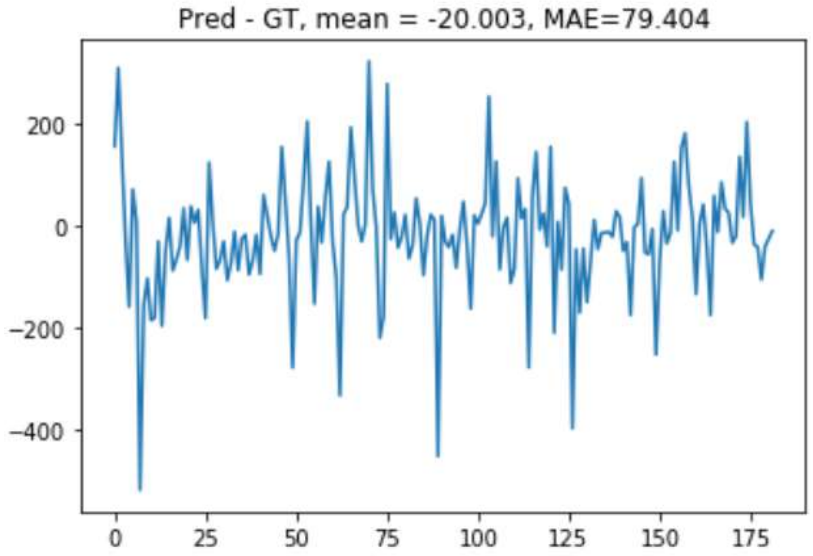


RESULTS OBTAINED

IN THIS IMAGE WE HAVE TRIED TO SHOW THE ACTUAL ERROR BETWEEN GROUND TRUTH AND PREDICTED VALUE WHERE WE CALCULATED THE DIFFERENCE BETWEEN THE GROUND TRUTH AND OUT PREDICTED VALUE OF COUNT AND PLOTTED ON THE GRAPH THAT CAN BE SEEN.

CODE:

```
error = np.array(ct_preds) - np.array(ct_gts)

plt.plot(error)

plt.title('Pred - GT, mean = {}, MAE={}'.format(

    str(round(np.mean(error), 3)),

    str(round(np.mean(np.abs(error)), 3))

))

plt.show()
```



RESULTS OBTAINED

1)THIS IMAGE IS OF FEW OF THE WORST CASE SAMPLES OUT OF ALL THE IMAGES THAT HAVE WORKED ON.

2)THE FORMULA FOR MEAN ABSOULUTE ERROR-

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - y_i'|$$

3)THE FORMULA FOR ROOT MEAN SQUARE ERROR-

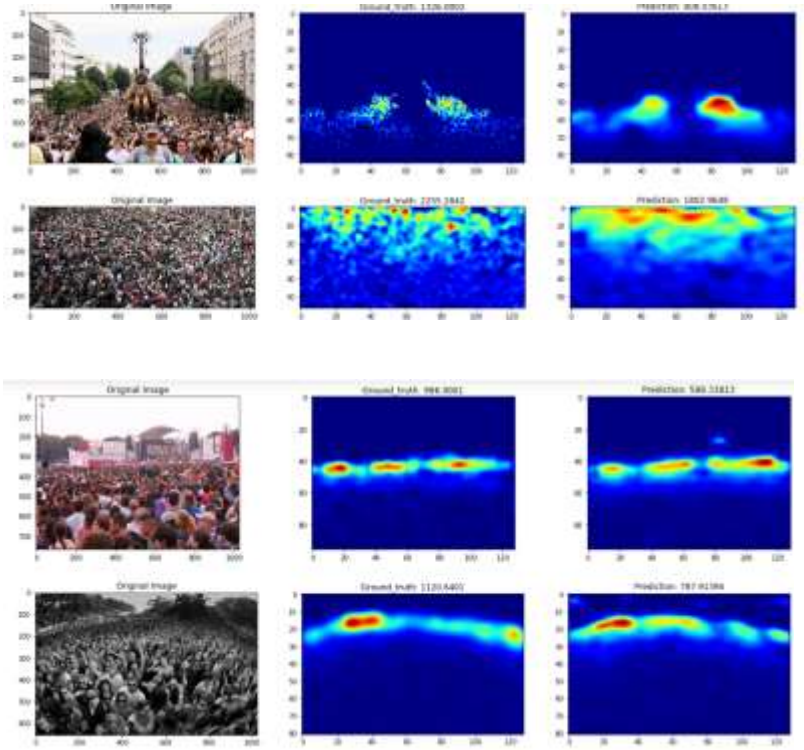
$$MSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - y_i')^2}$$

4)OUR PREDICTED VALUES, MAE – 79.4

RMSE – 117.36

5)THE VALUES IN THE PAPER, MAE – 58.6

RMSE – 97.8



CONCLUSION AND REFERENCES

CONCLUSION -

In this work, we proposed a novel architecture called a context-aware scale aggregation CNN-based crowd counting method (CASA-Crowd) that is trained in an end-to-end manner. Due to strong feature extraction property of deep neural network, we used deeper and wider networks to extract the deep and scale varying features. Furthermore, a dilated convolution approach is included in inception module to obtained the context-information. The performance of CASA-Crowd is comparable to the state-the-art methods due to varying receptive field and strong ability of handling the perspective varying issues. Moreover, the quality of density map is enhanced due to expanded spatial sampling. In this way, our proposed approach is capable of learning the low to complex, deeper and scale-aware features with enhanced density map.

REFERENCES -

1) A. Marana, L. D. F. Costa, R. Lotufo, and S. Velastin, ‘‘On the efficacy of texture analysis for crowd monitoring,’’ in Proc. IEEE Int. Symp. Comput. Graph., Image Process., Vis. (SIBGRAPI), Oct. 1998, pp. 354–361.

2) Y. Bharti, R. Saharan, and A. Saxena, ‘‘Counting the number of people in crowd as a part of automatic crowd monitoring: A combined approach,’’ in Information and Communication Technology for Intelligent Systems. Singapore: Springer, 2019, pp. 545–552.

----------THANK YOU-----------