Object Detection by a Super-Resolution Method and a Convolutional Neural Networks

Bokyoon Na
Dept. of Computer Engineering
Korea Polytechnic University
Siheungsi, Korea
bkna@kpu.ac.kr

Geoffrey C Fox
School of Informatics, Computing, and Engineering
Indiana University
Bloomington, USA
gcf@indiana.edu

***Abstract*—Recently with many blurless or slightly blurred images, convolutional neural networks classify objects with around 90 percent classification rates, even if there are variable sized images. However, small object regions or cropping of images make object detection or classification difficult and decreases the detection rates. In many methods related to convolutional neural network (CNN), Bilinear or Bicubic algorithms are popularly used to interpolate region of interests. To overcome the limitations of these algorithms, we introduce a super-resolution method applied to the cropped regions or candidates, and this leads to improve recognition rates for object detection and classification. Large object candidates comparable in size of the full image have good results for object detections using many popular conventional methods. However, for smaller region candidates, using our super-resolution preprocessing and region candidates, allows a CNN to outperform conventional methods in the number of detected objects when tested on the VOC2007 and MSO datasets**

Keywords- object detection, super-resolution, convolution neural networks, CNN, deep learning, machine learning

I. INTRODUCTION

Since Krizhevsky et al. [1] introduced specifically designed CNN architectures, there have been many methods to increase the rate of object classification on convolution neural networks. [2], [3], [4], [5], [6], [7] have shown performances to be increased. Nowadays, with many blur-less or slightly blurred images, CNNs classify objects with around 90 percent classification rates.

Recently, there has been research focused on reducing the misdetection and detection failures on convolution neural networks with the help of generative adversarial network (GAN). Furthermore, GANs have been extended by using reinforcement learning, [8].

Frameworks such as Caffe and Tensorflow can help to design CNN models for any specific computer visioning. Also, from a computer infrastructure point of views, graphics processing unit (GPU), such as those from Nvidia substantially increase performance. These advances allow one to design systems for vision detection and classification designed to run in real time.

However, even as there have been improvements of convolutional neural networks in multiple ways, there are still misdetections and detection failures for object classifications. For example, randomly 100 images from [9] were selected and preprocessed at three times lower resolution than the original images to build cropped images. Then they were interpolated by Bilinear and Bicubic algorithms, and finally, they were tested by [3]. Fig. 1 shows the detection failure of the second person from the left side of the image. The algorithm misses the second person on the left side in Fig. 1. We will compare several different algorithms and propose an advanced method in later sections.

Alex Krizhevsky et al. in [1] described the number of categories and designed their network with 1,000 object classes from ImageNet. [2], [5] classified 22 object classes from PASCAL VOC 2007 [9]. There is a 200 class dataset in ILSVRC2013, ILSVRC2017 contains several datasets and especially, the ImageNet dataset contains 1000 classes and Omniglot contains 1623 classes. But research is often done with VOC2007 20+1 classes even though there are datasets with more than several hundred classes.

Recently, the most popular image size on CNN has been around 256x256. Spatial Pyramid Pooling network, Faster R-CNN, and ConvNet are examples testing the relevance of the image size. Advances in camera systems and the popularity of mobile devices cause consumers to demand higher resolution images. Moreover, in medical and geometrical satellite imaging systems the demands of object classifications are required in strong. In [11], [12], [13], they show that recurrent neural network model is capable of extracting information from an image by selecting a sequence of regions and processing by selected region at high resolution.

Figure 1. An example of object classification of "person" by CNN.

In practical applications, there are many low-quality image processing systems such as surveillance camera systems, car black-box systems, or even mobile phone cameras for taking pictures of long distance. For example, in surveillance systems, it may not easy to increase the capacities for better image qualities because of their storage capacities, dark conditions on camera image sensors, and night visioning. Also in car black box systems, moving vibrations and the requirements of low electric power consumption may cause bad image compressions or lack of fast zooming or focusing devices. Notably, mobile phone pictures are blurred or have small target objects according to the limit of the software zooming algorithms.

Thus, we will introduce our research to improve object classification rates with improvements through super-resolution algorithms and convolution neural networks.

II. RELATED WORKS

To support fast implementation in compute-intensive cases such as video processing, the number of classes in CNN is preferentially kept small. The number of classes to detect or classify objects is a significant factor in the design of systems. The number of neurons, even in a hidden layer, is required to increase to handle many classes, and it causes the system to slow down because of several hidden layers associated with heavy computations. Papers such as [2], [14] have shown new methods or new parameters related to lower layers of neural networks. The paper [14] introduced a new feature extraction algorithm for object recognition. The paper [2] also improved CNN "region proposals" technique. Their new methods speed up object detection allowing the image dataset to run in almost real time.

Keiming He et al. in [5] had the best results for training and testing with the maximum image size of 392 pixels because they had variable size image datasets from VOC 2007 and ImageNet. Also, they showed results indicating that scale matters in the classification processing. Thus, they suggested the spatial pyramid pooling model support different image sizes in convolution layers which work with various image sizes while the standard fully connected layer requires a fixed image size. With the Caltech101 image dataset, they found objects that had better performances among several scaled datasets. They noticed that this is mainly because the detected objects usually occupy large regions of the whole images. They evaluated cropped or warped images and got lower accuracy rates than the same model on the undistorted full image.

In [15] Generative Adversarial Networks (GANs) are described as generative models that use supervised learning to approximate an intractable cost function and can simulate many cost functions, including the one used for maximum likelihood. More specifically, a generative model g trained on training data X sampled from some true distribution D is one which, given some standard random distribution Z, produces a distribution D' which is close to D according to some closeness metric (a sample $z \in Z$ maps to a sample $g(z) \in D'$).

There are several points of researches related to the GANs which are training auto-encoder based GANs [4], learning semi-supervised and generating images that humans find visually realistic [16], and training for semi-supervised text classification [17]. Also, there are GAN extensions related to super-resolution methods and reinforcement learning.

Image super-resolution (SR) means a deep learning method which infers a high-resolution image from a low-resolution image. In [18] two types of SR algorithms are introduced: single image based and multiple images based. In the multiple image SR algorithm, two low-resolution images of the same scene are fed in as input and a registration algorithm to find the transformation between them is added. The single image SR algorithm employs a training step to learn the relationship between a set of high-resolution image and their low-resolution counterparts and this relationship is used to predict the missing high-resolution details of input images. They can be applied for video data to improve action recognition.

Using single image SR, [19], [20], and [21] present methods using a mixture of experts (MoE) which are anchor based local learning approach, sparse coding, and deep convolution neural networks, respectively. Christian Ledig et al. in [22] proposed a GAN method using generator networks and discriminator networks to recover photorealistic natural images with minimization of the mean squared reconstruction error.

In SR and CNN algorithms, image interpolations are used to resize images or to crop and warp image patches. In particular, a small region of interest in a large image is processed with cropping or warping to fit into bounding boxes or region proposal, and it results in a decreased object detection rate.

Current CNNs, which are using the sliding window method, can process variable size images as their input. However, the fully connected layer can work only with a fixed size of feature maps. For this reason, CNNs initially used a fixed size for input images. Kaiming He et al. [5] introduced a variable size of images as the input of their CNNs and warped or cropped images to get fixed feature maps. Cropped or warped images might have poor object recognition results. This caused us to implement super-resolution instead of interpolation methods.

III. PROPOSED IMAGE RESOLUTION IMPROVEMENTS

CNN requires a target image to be a fixed size before processing hidden layers. Thus with a given input image, the network resizes it to the target size of the image. In [2], the bilinear interpolation algorithm is applied to extend the image size to be an appropriate input for the network. As future research, the extended image size directly adopted by the super-resolution method will be left. Instead of this approach, we propose a super-resolution method to increase the image size to around 492x324 pixels.

Thus, we will describe super-resolution methods first to improve the image quality in pre-processing before the input layer of CNN; then, we describe how the preprocessed image can be classified by a CNN.

A. Super-resolution image scale-up

We know there, in image analysis, have been improvements for object recognition and there is research on better recognition rates even though one often only gets a small improvement. Also, this work largely focuses on deep hidden neural networks. However, in this section, we will focus on the pre-processing of image samples which are inadequate rather than on the improving of the hidden layers. For example, if a cropped image whose size of 166x110 is extracted from a normal image, it is too small to feed into the input layer of our CNN.

As an expert which means that each of the model component classifiers or regressors is highly trained, component regressors, W_i, and as an anchor point, v_j have relations such as:

$$\min_{\{v_1,v_2,,\ldots,v_k;\, w_1,w_2,\ldots,W_k\}} \Sigma_{j=1}^{N} \sum_{i=1}^{K} c_{ji} \| h_j - W_i l_j \|^2,$$

where l_j means a low resolution path, h_j is the corresponding high resolution patch, v_j is the nearest anchor point for l_j and c_{ji} is a continuous scalar value which represents the degree of membership of l_j.

However, to improve the computational efficiency and the competitive image quality of anchor-based local learning method of multiple regressors, a mixture of experts which is one of the conditional combined mixture models [23, 24] is proposed here. As in [24] we define the model of mixture of experts for super-resolution images, describing the expectation and maximization (EM) algorithm, and also train and test this model.

As a mixture of experts model, a maximum likelihood estimation should be solved iteratively by the EM algorithm generally. Every iteration the posterior probabilities are calculated for patches, and then we get the expectation of the log-likelihood as an E-step. During M-step, anchor points and regressors are updated which is a softmax regression problem. After training, super-resolution images can be constructed by collecting all the patches from regressed low-resolution patches and averaging the overlapped pixels.

To compare the performance of the super-resolution method with interpolation methods, bilinear and bicubic interpolated images will be built in addition to the images produced by the super-resolution method.

B. Object detection

We implemented the convolution neural network based on the Faster R-CNN [3] because it supports variable image sizes as the input data of CNN and sliding widow proposal scanning for the convolution network. Above all, the detection speed is fast enough like almost real-time. As mentioned earlier, Faster R-CNN uses region proposal to detect an object. But the size ratio of region proposal to the whole image is critical the object to be detected. It means we can distinguish our proposed method compared to other interpolated data.

The CNN is implemented based on Intel® Xeon CPU 2.30GHz and 4 NVIDIA Tesla K80 GPU boards. Each GPU board has a memory of 12GB and two GPUs.

Our CNN has several convolutional layers to support a region proposal network in addition to the conventional CNN [1], [7]. Also, these convolution layers are shared with object detection networks as in [5]. Thus, our model works as a deep CNN which has more convolution and pooling layers too.

As multiple-scale prediction schemes for regression references, there are schemes based on multiple-images, on multiple-filters, and on multiple-anchors in [3]. With multi-image schemes, images are resized at multiple scales and feature maps are computed for each scale. Even though this scheme is time-consuming, our super-resolution method can resize images to get better prediction scores. However, for less usage of computational resources, for fast detections, and for feature sharing in fully convolution layer, we choose the the multi-anchors scheme.

As additional, for the memory usages of CAFFE, we have constraints on the number of images at reading input images from image dataset. In our deep CNN model allowing multiple scaled image sizes, the number of region proposals is w (width of the image) x h (height of the image) x k (the number of anchors of a region proposal). It means that memory space for the only region proposal network is simply required severely. Therefore, we constraint the number of images in hidden layers less than or equal to 20 images simultaneously.

For model training, we use pre-trained model parameters taken from Faster R-CNN implementation. Faster R-CNN had training, validating and testing with PASCAL VOC 2007 of 9000 images. As a result, we consider that the parameters are fitted well enough to our model.

For each proposal from an image, an object is roughly considered, and the proposal size is selected one of predefined window size. Thus, this patch of an image may be interpolated. If instead of interpolation methods the super-resolution method is adopted, any variable sizes of input images with variable sizes or shapes can be supported with better cropped or warped regions.

We will not use the super-resolution method to improve the patch image qualities here. We keep this for future research.

IV. PERFORMANCE EVALUATIONS

Before any other processing, we considered several image file formats to get better image quality for pre-processing, training, and testing images. Therefore, we picked two image data formats and with a few images we went through the full cycle of our proposed model. While most of the image datasets are built images based on the JPG file format, this did not seem to offer as good results in super-resolution

processing compared with the BMP format. As the result of this simple testing, we decided to convert images from a file format of JPG into BMP format as a pre-processing procedure. Then, the image is taken to be further processed with bilinear, and bicubic interpolations, and the super-resolution method. Also, we scale down image size by 3 times smaller in each width and height to build images with lower quality, instead of collecting images by cropping or warping images.

In our model, we implement the super-resolution procedure based on [19]. Rather than train with less than 50 images as in most of the published super-resolution models, we have trained our model based on their initial parameters and with 100 PASCAL VOC2007 images. There is no overfitting with this number of images for training. With random images from PASCAL VOC2007 [9], [25] and test images from [26], we have tested our super-resolution model. We could not find any differences between them. Therefore, we will test the object classification with 520 images randomly extracted from VOC2007 and 1224 images from MSO [27].

In **Error! Reference source not found.**, we found the dataset has a different number of objects compared to our intuition. For example, the first image of Fig. 2 is labeled with no object, but our proposed model detected an object or objects, like as shown in the image. Thus, we decide not to use the given label from the dataset.

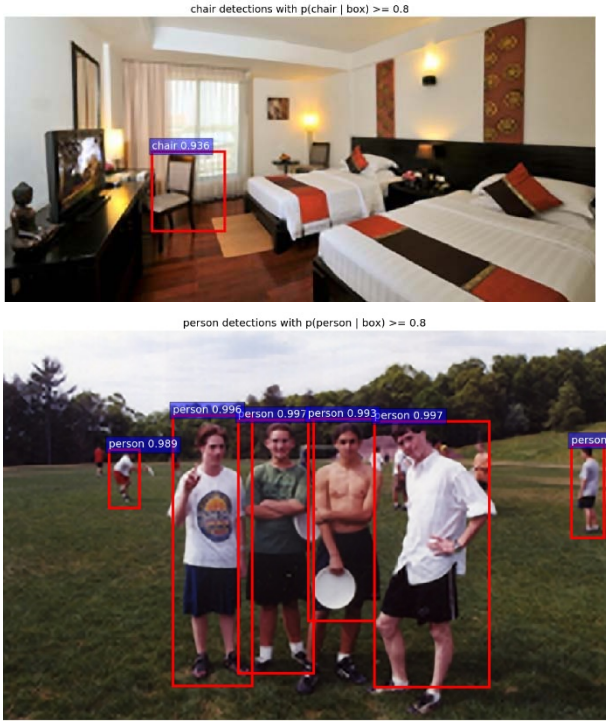

Figure 2. Images which have objects detected by our method but are labeled with no objects or a lesser number than we find.

A. *Comparison of Output Pictures*

Many images from the PASCAL VOC 2007 have multiple objects, as shown in the results given in **Error! Reference source not found.**with 3 different pre-processing models, which are a bilinear interpolation, bicubic interpolation, and super-resolution. Our model is set with a learning rate of 0.001 and detection scores with 0.8 or higher.

The first column shows the number of detected objects in each class as given by bilinear interpolation. For the second column, we show for bicubic interpolation, the number of objects detected and the detection rate. The third column shows analogous results for our super-resolution approach.

As shown in **Error! Reference source not found.**, our model detects more objects than the other two models, but the average detection scores are not very different. It means that our model makes improved input images, which are fed into the CNN which is able to increase the number of detections. **Error! Reference source not found.** shows the cumulative number of detected objects and compare them.

With image dataset of Microsoft MSO, we present results in **Error! Reference source not found.** and **Error! Reference source not found.**4. As with the VOC2007 dataset, our proposed model detects more objects than the two other models. Unlike the case of the VOC2007, the MSO dataset has images with no objects. However, the average detection scores here for our model is similar to that VOC2007.

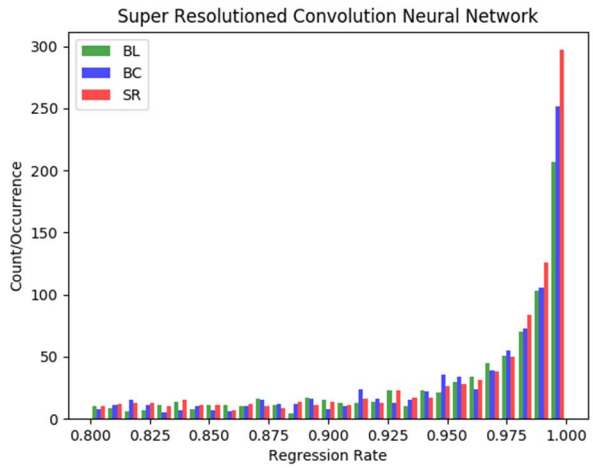

Figure 3. Histogram for convolution network models with 3 different image pre-processing methods on randomly extracted images from VOC 2007 dataset

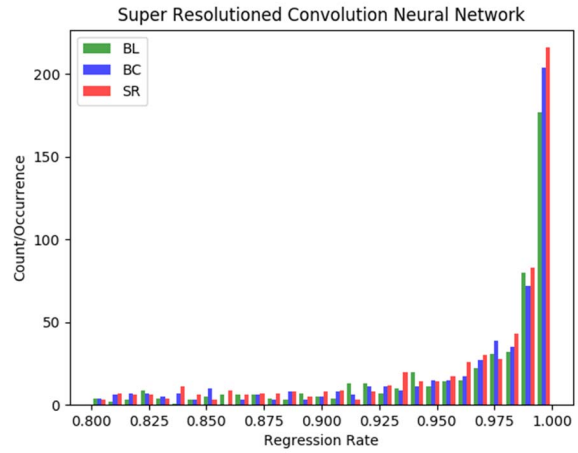

Figure 4. Histogram for convolution network models with 3 different image pre-processing methods on randomly extracted images from MSO dataset

TABLE I. DETECTED OBJECTS ON 3 DIFFERENT PRE-PROCESSING AND CONVOLUTION NEURAL NETWORKS WITH PASCAL VOC2007 DATASET. #TP IS THE NUMBER OF TRUE POSITIVES CORRECTLY PREDICTED. THE COLUMN LABELED MEAN IS THE AVERAGE PROBABILITY FROM THE METHOD.

	Bilinear		Bicubic		Ours	
Classes	#tp	mean	#tp	mean	#tp	mean
aeroplane	25	0.9569	29	0.9472	27	0.9787
bicycle	16	0.9602	17	0.9697	18	0.9520
bird	18	0.9204	25	0.9178	30	0.9500
boat	16	0.9330	18	0.9260	20	0.9276
bottle	19	0.9222	17	0.9208	15	0.9331
bus	26	0.9626	25	0.9639	26	0.9588
car	93	0.9662	100	0.9671	104	0.9710
cat	11	0.9427	10	0.9665	11	0.9739
chair	25	0.9273	34	0.9368	45	0.9320
cow	11	0.9149	12	0.9216	16	0.9385
dining table	7	0.9317	7	0.9420	12	0.9193
dog	37	0.9559	36	0.9657	35	0.9565
horse	30	0.9560	34	0.9574	37	0.9694
motorbike	13	0.9467	13	0.9630	16	0.9581
person	405	0.9546	432	0.9575	466	0.9590
potted plant	10	0.9467	12	0.9194	18	0.8767
sheep	15	0.9275	13	0.9380	14	0.9227
sofa	7	0.9191	8	0.9175	8	0.9475
train	7	0.9193	4	0.9276	4	0.9572
tvmonitor	25	0.9653	25	0.9729	26	0.9479
Total	816	0.9415	871	0.9450	**948**	**0.9465**
misclassified	25		21		**39**	

TABLE II. DETECTED OBJECTS ON 3 DIFFERENT PRE-PROCESSING AND CONVOLUTION NEURAL NETWORKS WITH MSO DATASET. #TP IS NUMBER OF TRUE POSITIVES CORRECTLY PREDICTED. THE COLUMN LABELED MEAN IS THE AVERAGE PROBABILITY FROM THE METHOD.

	Bilinear		Bicubic		Ours	
Classes	#tp	mean	#tp	mean	#tp	mean
aeroplane	5	0.9650	6	0.9300	9	0.9263
bicycle	1	0.9992	2	0.9112	1	0.9978
bird	50	0.9512	59	0.9521	75	0.9627
boat	1	0.8301	1	0.9771	1	0.9713
bottle	23	0.9074	24	0.9062	21	0.9117
bus	3	0.9954	3	0.9954	3	0.9871
car	16	0.9641	19	0.9558	18	0.9499
cat	11	0.9639	10	0.9829	11	0.9540
chair	19	0.9299	20	0.9270	22	0.9393
cow	5	0.9452	6	0.9488	7	0.9584
dining table	3	0.9410	4	0.8927	4	0.9072
dog	42	0.9559	47	0.9636	50	0.9478
horse	11	0.9728	11	0.9726	15	0.9362
motorbike	4	0.9487	4	0.9529	4	0.9588
person	302	0.9715	318	0.9718	340	0.9719
potted plant	4	0.9023	5	0.8756	9	0.9035
sheep	1	0.8780	1	0.9353	4	0.9064
sofa	3	0.9204	3	0.9099	6	0.9261
train	6	0.9746	7	0.9529	8	0.9522
tvmonitor	6	0.9720	6	0.9804	10	0.9248
Total	516	0.9444	556	0.9447	**618**	**0.9447**
misclassified	86		82		**92**	

B. Big vs. Small ROI Pictures and Their Detection Rates

As mentioned previous section, if objects are big enough compared to the size of the image which is containing the object proposal, objects from interpolated images with bilinear or bicubic methods are identified satisfactorily and sometimes may have better performance than our proposed model. However, our proposed model has much better results. with objects from small bounding boxes or small ratio of object size to the size of the image which contains the object. In **Error! Reference source not found.**5, our proposed model detects a chair which is a quite small object in the image, while the other two models did not detect this 'chair' object. Even though the other chair is detected in all of three models, our proposed model has a little bit higher score.

V. CONCLUSION

We proposed a model which consisting of super-resolution processing and a convolution neural network. In this model, several kinds of classes from two different datasets are scored when objects are detected. Our model has improved the number of objects detected compared with other state of the art convolution neural networks which use bilinear interpolation algorithms for interesting regions. Our improvement is especially clear in the case of small object detection.

Figure 5. Comparison of small object detection through Bilinear, Bicubic, and SR models

In this proposal, we did not implement the restriction that only bounding box areas are processed with the super-resolution method. But we expect that this can save the computational resources and thus we can adopt this in real-time processing for better object detection or classification.

Our proposed model appears powerful in scenarios such as relatively small objects in big pictures, warping in the region proposals approach, and object detection from cropped images.

VI. ACKNOWLEDGEMENT

This work was supported by the 2017 sabbatical year research grant of the Korea Polytechnic University.

REFERENCES

[1] Alex Krizhevsky, Ilya Sutskever, Geoffrey E. Hinton. "ImageNet Classification with Deep Convolution Neural Networks". In Advances in Neural Information Processing Systems (NIPS), (1097-1105). (2012).

[2] Girshick Ross. "Fast R-CNN". arXiv: 1504.08083v2 [cs.CV] 27 Sep 2015.

[3] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. "Faster R-CNN: Towards Real-time Object Detection with Region Proposal Networks". arXiv:1506.01497v3[cs.CV] 6 Jan 2016.

[4] David Berthelot Schumm, Luke Metz Thomas. "BEGAN: Boundary Equilibrium Generative Adversarial Networks". arXiv: 1703.10717v2 [cs.LG]. (2017).

[5] Keiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. "Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recogniton". arXiv:1406.4729v4. (2015).

[6] Ross Girshick, Jeff Donahua, Trevor Darrell, Jitendra Malik, "Region-Based Convolution Networks for Accurate Object Detection and Segmentation", *IEEE Transactions on Pattern Analysis and Machine Intelligence Vol. 38, NO.1*, 2016.

[7] Karen Simonyan, Zisserman Andrew. "Very Deep Convolutional Networks for Large-Scale Image Recognition". ICLR. (2015).

[8] Ian J. Goodfellow Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courvill, Yoshua BengioJean. "Generative Adversarial Nets". arXiv:1406.2661v1. (2014).

[9] M. Everingham Van Gool, C. K. I. Williams, J. Winn, and A. ZissermanL. "The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results". (2007).

[10] Achanta R., Estrada F., Wils P., Süsstrunk S. "Salient Region Detection and Segmentation". In: Gasteratos A., Vincze M., Tsotsos J.K. (eds) Computer Vision Systems. ICVS 2008. Lecture Notes in Computer Science, vol 5008. Springer, Berlin, Heidelberg. (2008)

[11] Volodymyr Mnih, Nicolas Heess, Alex Graves, and Koray Kavukcuoglu. "Recurrent Models of Visual Attention". arXiv:1406.6247v1[cs.LG] 24 Jun 2014.

[12] Jimmy Lei Ba, Volodymyr Mnih, and Koray Kavukcuoglu. "Multiple Object Recognition with Visual Attention". arXiv:1412.7755v2[cs.LG] 23 Apr 2015.

[13] Karol Gregor Danihelka, Alex Graves, Danilo Jimenez Rezende, Daan Wierstra Ivo. "DRAW: A Recurrent Neural Network For Image Generation". arXiv:1502.04623v2[cs.CV] 20 May 2015.

[14] J.R.R. Uijlingsvan de Sande, T. Gevers, and A.W.M. Smeulders K.E.A. "Selective Search for Object Recognition". IJCV. (2012).

[15] Ghosh Nachum and Debiprasad Ofir. https://www.quora.com. "What-are-Generative-Adversarial-Networks-GANs".

[16] Tim Salimans Goodfellow, Wojciech Zaremba, Vicki Cheung Ian. "Improved Techniques for Training GANs". arXiv:1606.03498v1. (2016).

[17] Takeru Miyato M Dai, Ian Goodfellow Andrew. "Adversarial Training Methods for Semi-Supervised Text Classification". ICLR. (2017).

[18] Nasrollahi Kamal, Guerrero Escalera Sergio, Rasti Pejman, Anbarjafari Gholamerza, Baro Xavier, J. Escalante Hugo, Moeslund B. Thomas. "Deep Learning based Super-Resolution for Improved Action Recognition". In International Conference on Image Processing Theory, Tools and Applications (IPTA) IEEE Signal Processing Society. (2015).

[19] Kai Zhang Wang, Wangmeng Zuo, Hongzhi Zhang, Lei ZhangBaoquan. "Joint Learning of Multiple Regressors for Single Image Super Resolution". IEEE Singnal Processing Letters. Vol. 23 No. 1. (2016).

[20] Zhaowen Wang Liu, Jianchao Yang, Wei Han, Thomas Huang Ding. "Deep Networks for Image Super Resolution with Sparse Prior". ICCV. (2015).

[21] Chao Dong Change Loy, Kaiming He, Xiaoou Tang Chen. "Image Super Resolution Using Deep Convolution Networks". arXiv:1501.00092v3.v (2015).

[22] Christian Ledig Theis, Ferenc Huszar, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, Wenzhe ShiLucas. "Photo-Realistic Single Image Super Resolution Using a Generative Adversarial Network". arXiv:1609.04802v5[cs.CV]. (2017).

[23] Bishop Christopher. "Pattern Recognition and Machine Learning". Springer. (2006).

[24] Zhang Zhang, Baoquan Wang, Wangmeng Zuo, Hongzhi ZhangKai. "Joint Learning of Multiple Regressors for Single Image Super-Resolution". IEEE Signal processing letters, Vol. 23, No.1. (2016).

[25] Olga Russakovsky Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg and Li Fei-Fei Jia. "ImageNet Large Scale Visual Recognition Challenge". arXiv:1409.0575. (2014).

[26] Olga Russakovsky Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg and Li Fei-Fei Jia, "ImageNet Large Scale Visual Recognition Challenge 2017". (2017).

[27] Zhang Ma, Shuga Sameki, Mehrnoosh Sclaroff, Stan Betke, Margrit Lin, Zhe Shen, Xiaohui Price, Brian Much, Radom Jianming. "Salient Object Subitizing". IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2015).