International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878 (Online), Volume-8 Issue-6, March 2020

Improving Efficiency of CNN using Octave Convolution

A.V.Sriharsha, K.Yochana

Abstract: ***In recent years, the Convolutional neural networks (CNN) has been active in various Artificial intelligence applications as well as computer vision tasks. We suggested an effective technique in this study to decrease the number of duplicates in feature maps of CNN. Proposed a novel convolution scheme Octave convolution (Octconv) to minimize the duplicates in the feature maps and boost the CNNs performance. The principle concept of this method is to separate the Convolutional filters into a higher frequency and lower frequency sections. In this report, we made an attempt for minimizing the spatial redundancy directly from output feature maps of CNN using the following 3 steps: First, divide the channels into higher and lower frequency parts depending on the information of the image using Multi-scale representation. Second, reduce the number of FLOPs from the low frequencies. Third, before sending the output to combine both the higher frequency and lower frequency information of the image. The key purpose of this abstract is to improve CNNs efficiency by reducing spatial redundancy in the feature maps of the convolution layer.***

Keywords: Octave Convolution, Convolutional Neural Networks, Multi-scale Representation, Feature maps.

I. INTRODUCTION

In the last couple of years, CNN have been achieved significant outcomes in different computer vision and artificial intelligence some applications like object recognition, object detection and image retrieval. With the recent efforts to decrease the inherent redundancy of dense model parameters [1,2]within the channel dimension of feature maps[3,4], the efficiency of CNN is continuously increased. The general trend to improve performance further has made models more complex and deeper. A CNNs simple architectural design begins processing on a higher-resolution input, in which filters analyse lower local sections. Recent attempts have made improvements to the Convolutional layers by minimizing their inherent redundancy in dense model parameters and also in the channel dimension of standard Convolutional layers are the main element on such architecture. These have been suitable for local identifications connections with features from the earlier layer which are always the same spatial resolution and map their appearance to a feature map. Real images can be splitted into a lower frequency signal capturing the global layout and also the coarse structure as well as higher frequency component capturing the information. Previous researches focuses on developing new network architectures and the refinement of the CNN models. Many research frameworks also focus on the development of new non-linear functions, including Rectified Linear Unit (ReLU).

Increase accuracy by increasing complexity of the model with a deeper network is not for free: the cost of calculation increases enormously (FLOPs). Several types of convolution filters were therefore proposed to reduce the FLOPs and increase the models efficiency. Existing Convolutional filters can be divided approximately into two groups: 1- Depth-wise Convolution Filter [5] to perform Depth-wise convolution (DWC) 2- Group-wise Convolutional [6] Filter to perform group wise convolution (GWC). In the most recent architectures, the model is effective using a combination of these filters. Many of the popular models have been using these convolutions to explore new FLOPs architecture.

A natural picture might be splitted into a lower frequency component which describes the slowly changing structure and a higher frequency component that describes the fine information that are rapidly changing [7,8]. Similarly, the output feature maps of different spatial frequencies and proposes a new representation of multi-frequency function that saves higher frequency and lower-frequency feature maps into several groups. We are using vanilla convolution instead of a substitution of vanilla convolution with octave convolution (Octconv). Octconv uses significantly fewer memory and computing resources in the substitution of vanilla convolution

Although a study has illustrated that the network has a no.of parameters and it has the no.of accurate results it produces, various analysis were performed to improve accuracy by increasing the inference speed and decreasing the model size. Octave convolution (Octconv) is a way to minimize the spatial redundancy. The feature maps are splitted into higher and lower frequencies. The higher frequency contains the fine information and the lower frequency contains the rough information. Lower spatial resolution was used for the lower frequencies to minimize the spatial redundancy, and both the memory consumption and computational cost are minimized successfully without reducing the precision.

Manuscript received on March 15, 2020.
Revised Manuscript received on March 24, 2020.
Manuscript published on March 30, 2020.
* Correspondence Author
Dr. A. V. Sriharsha*, Professor, department of Computer Science and Engineering, Sree vidyanikethan Engineering College, Tirupati, India. Email: avsriharsha@vidyanikethan.edu
Ms. K. Yochana, PG Scholar, department of Computer Science and Engineering, Sree vidyanikethan Engineering College, Tirupati, India, Email: kakarlyochana009@gmail.com

In this paper, new approach Octave convolution (Octconv) was introduced. Our motivation is that to decrease the volume of cost resolution to tackle memory consumption and the computational cost. However, the only cause of performance degradation can reduce the resolution. Octconv focuses on cost volume of spatial redundancy. The features of the image are divided into higher and lower spatial features, with less resolution. Both the cost volumes in the proposed Octave convolution are developed by the respective features.

We therefore, primarily split the image into two sets: a higher frequency component with fine-structure information and a lower frequency part with the rough-structure information. A lower resolution is applied to reduce the spatial redundancy of the lower frequency costs volume, which can the reduced memory consumption and computational cost. A new model was developed to achieves an octave convolution.

II. RELATED WORK

The literature survey is mainly done by taking deep insight into different literature sources and the following literature reviews is focus on different parameters and methods that are used in minimizing the spatial redundancy in convolutional neural networks. In comparision with standard convolution operations HetConv (Heterogenous Kernel-Based Convolution) [9] minimizes the computation and the no.of parameters while still retaining the representational efficiency. Compared HetConv with depth-wise, Group-Wise point-wise and standard convolutions on different architectures. HetConv convolution is more powerful than existing convolutions. Proposed a novel filter (HetConv) that contains a heterogeneous kernel to reduce the FLOPs of existing model. HetConv performed several experiments with the current state-of-art architectures. Extracted images from CIFAR [10] datasets. Using ResNet-34, ResNet-50 [11] and VGG-16 architectures [12], three large scale experiments were carried out with Imagenet.

CNN architecture for the study multi-scale representations [13] was proposed based on integrating several networks with strong interactions between speed and precision in object and speech recognition. Network pruning and quantizing are common for eliminating model redundancy and reducing computational cost. Validated the efficiency of bL-Net tasks of object and speech recognition. Some CNN models allows bL-Net to be easily integrated. Used ImageNet dataset for object recognition. When operating at low inputs, the big branches achieve a substantial computational reduction, while the small branches incorporate features from high inputs with light computation. bL-Net shows the FLOPs which are reduced will increase the run time on the GPU consistently.

Octave deep plane- sweeping network (OctDPsNet) [14] plane-sweeping stereo to minimize the spatial redundancy of a plane-sweeping stereo for learning based. The lower and higher frequency spatial frequency features are used to produce two cost volumes one at a lower resolution by using plane sweeping approach. The images are extracted from SUN3D [15], RGB-D SLAM [16], MVS [17] and scenes11 [18] for training and evaluation. These include photos, depths and camera locations, as well as various outdoor and indoor scenes. The datasets were distributed into testing and training datasets. OctDPSNet has implemented different methods on five datasets and has decreased memory and processing time.

Proposed an effective method for stabilizing the training of GANs (Generative Adversarial Networks) [18] based on CNN. Convolution scheme for stabilizing the training and reducing the chance of mode failure. OC-GAN generalizable scheme for GANs leading to make training more stable. Trained OC-GAN on CelebFaces [19] Attributes (CelebA) dataset. Tackled the problem of stability during GANs training. Due to the octave convolution implementation, the OC-GAN is coined as a Simple framework. This approach complements existing stabilization methods and it is orthogonal and can be connected simply to any CNN based GAN architecture.

Reduces parameter redundancy using sparse decomposition [21]. Maximum sparsity is achieved by utilization with maximum sparsity of both of the inter and intra channel redundancies. Provided an effective CPU Sparse Convolutional Neural Networks (SCNN) model for the sparse matrix multiplication algorithm. In the SCNN model, a few convolutional layers can be done with each sparse convolutional layer followed by a sparse matrix multiplication. Trained the model on ImageNet LSVRC [22] 2012 dataset. The model consists of five layers of the convolution and two layers of fully connected. The working time of the fully connected layers are less than the convolutional layers.

Reduces the spatial and channel redundancy directly from the visual input for CNNs acceleration. ESPACE (Elimination of Spatial and Channel Redundancy) [24] used to decrease the consistency of 3D channels in convolutional layers through low-level approximity of convolutional filters. Access the process of convolution by avoiding unsalient visual component spatial positions. Accelerated network is trained using training data through Back-propagation. 1Evaluated on the dataset of ImageNet 2012. Here each image is associated with one ground truth category. Trained ESPACE model on evaluating group of ImageNet 2012, and evaluate it on validation set using central path only.

A number of works were performed to compact a deep network to decrease memory size and possibly computation [25]. A generic structure of a neural network with multi-linear projection was proposed. When compared with traditional CNN it takes many times less memory through inheriting similar CNN design principles. CIFAR-100 [26], SVHN [27] and Tiny-ImageNet [28] datasets are used for extracting images. The difficulty can be controlled by the resulting structure with flexibility by changing the no.of projections in every mode by hyper parameter R.

The existence of redundant neurons of neural network filter is investigated [29] in networks this phenomenon is increased with filter numbers in one layer. Observes the occurrence of duplicate filters during training iterations, investigates its concentration factors and compare study the factors that affect their concentration and compare existing reduction network operations. Used two simple network architectures a fully connected MLP and a CNN. Filter is used to denote both the channels of a convolutional layer and also indicate an individual neuron's weights of a fully connected layer in the weight matrix. Images are extracted using CIFAR-10 dataset.

In MLPs, filter duplication happens better than the CNNs, and it appears that these results are extracted from over-parameterization in the fully-connected MLP model. More number of filters are applied to a layer and more duplicates in MLP are possible and less are identified in CNNs

Accelerate and compress deep neural networks to use CNN models in low devices such as mobile phones or built-in gadgets [30]. Concentrated on pruning at the filter level if it has less importance then the whole filter is removed. To further minimize pruned model dimensions, "gcos" (group convolution with filters), proposed a more detailed convolution with filters. The ILSCVR-12 VGG-16 and ResNet-50 pruned by two widely used networks. The gcos are then added to our frame. This part introduces two small models produced through our Thinet. Then many applications of Thinet were introduced. The default run count is 1. In CAFFE Thinet, an effective method of channel-wise pruning for deep model acceleration and compression, all the experiments are conducted. The suggested scheme will substantially increase the performance of models with respect to existing methods. In addition, our models can be paired with any existing compression models.

Convolutional neural networks (CNNs) revolutionize machine learning, but they present significant programming challenges [31]. A variety of FPGA based accelerators have suggested increasing CNN's performance and efficiency. Current methods are designed to create a single processor that measures the CNN layers at a time and the processor is designed to optimize the calculation of the layer array. The CNN accelerator has been evaluated by using our method for four networks: AlexNet, VGGNet-E, SqeezeNet and GoogleNet. This method is designed to address two networks: Xilinx virtex-7 FPGAs (485T and 690T). Considered designs of both the single precision floating-point and also 16-bit fixed point arithmetic. The optimization algorithm was developed to provide efficient designs for MultiCLP with in a limited resource budget (DSP slices, BRAMs, and bandwidth).

Proposed a new method to reduce the computational costs of testing neural networks that have limited their implementation in low-energy devices like mobile phones [32].The implementation of three CNNs with increasing size and device complexity begins with comparisons between the proposed perforation masks in the same benchmark set acceleration of a single AlexNet layer. In this sense experiments are promoted by comparisons of the proposed perforation masks. The Caffe implementation is used for AexNet, which is different from the original architecture. A MatConvNet framework is used to complete all experiments, except for AlexNet and VGG-16 fine tuning, for which we use a Caffe fork. Perforated CNNs achieve lower error compared to the baseline, are more stable and do not modify the architecture of a CNN

Handwritten digit recognition is an extensively well investigated area where pre-segmented handwritten digits can be distinguished [34]. CNN is highly costly that is generally used for ensuring high accuracy in complex classification problems that require the tuning of millions of parameters. The CNN has a kernel size of 5×5 with zero padding=2 and stride= (1,1) in both the first and also the second convolutional layers. . As an input of the first fully connected layer, then the second convolution and also the max pooling layer, in which each feature map comprising a maximum of 7*7 = 49 pixels are generated. The experiments are conducted by varying the feature sizes on the MNIST datasets. Minimized the space used to evaluate the model and compare the precision, test precision and execution time of the problems trained with CNNs using the entire feature

S.no	Title	Year	Method	Datasets	Remarks	Ref.no
1	HetConv: Heterogeneous kernel Based Convolutions for Deep CNNs	2019	CNN, VGG, Residual network	CIFAR-10	Convolutional does not increase the number of layers to get FLOPs reduction so latency is zero	[9]
2	An effective Multi scale feature representation for visual and speech recognition	2018	Deep Convolutional Neural Networks, Multi scale networks	ImageNet, Switchboard	The computational complexity in multi scale networks has not been addressed much	[13]
3	Octave Deep Plane Sweeping Network: Reducing Spatial redundancy for learning-Based Plane-sweeping	2019	Convolutional Neural Networks	SUN3D, SLAM, MVS and scenes for training and evaluation. ETH3D Used for evaluation not for training	Improvement of the low frequency ratio α based on scene properties is expected	[14]
4	Stabilizing GANs with Octave Convolution	2019	CNN	Celeba	--	[18]
5	Sparse Convolutional Neural Networks	2015	Deep Neural Networks	ILSVRC 2012	Method is 2% lower than original spatial pyramid pooling while at the same time achieves many times faster speed	[21]
6	Accelerating CNN Via Eliminating Spatial and Channel Redundancy	2017	Back propagation, CNN	ImageNet 2012	lead to state-of-the-art rate distortion with less than 1% accuracy	[24]
7	Improving Efficiency in CNN using Multi linear filter	2018	CNN	CIFAR-10 ,CIFAR-100, SVHN and Tiny-ImageNet	Mapping includes two computing schemes which allow either a reduction in computation and memory when R is high, or scalability when R is small.	[25]
8	Reducing duplicates in Deep Neural Networks	2017	CNN,MLP	CIFAR	Duplication of filters in MLPs occurs more than in CNNs and this tends to be the consequence that the MLP model has been over parameterised	[29]
9	Pruning CNN filters for a Thinner Net	2019	CNN, ThiNet, VGG	CAFFE	---	[30]
10	Maximizing CNN Accelerator through Resource partitioning	2017	CNN	CIFAR	--	[31]
11	Acceleration through Elimination of Redundant Convolutions	2016	CNN, VGG-16	CAFFE	Perforated CNNs achieve lower error compared to the baseline, are more stable and do not modify the architecture of a CNN	[32]

12	Feature Map Reduction in CNN for Handwritten Digit Recognition	2019	CNN	MNIST	CNN is computationally cost which results in waste of resources for less challenging research issues.	[34]

Table I: Literature survey

III. METHODOLOGY

A. Overview

The overview of the approach we suggested is similar to previous CNN models. The major difference is we are processing higher and lower spatial frequency features separately, both for the image features and cost volume. Integrates Octconv into depth and group-wise convolution architectures. Octave Convolution directly operates on multi scale Representation. Here input featured maps factorized into High and low frequency groups. Reduces spatial redundancy in the CNN feature maps by down sampling/ compressing of low frequency maps spatial resolution. Divide convolutional features into two sets at various spatial frequencies and process them at their respective frequency with different convolutions one octave a part. Since the resolution can be reduced for the lower frequency, storage and computation are saved. Develop a plug-and-play operation called Octconv to directly substitute with the vanilla convolution and reduce spatial redundancies.

Octave Convolution can be rewritten as:

$$Y^H = f(X^H; W^{H\to H}) + upsample\,(f(X^L; W^{L\to H}), 2)$$
$$Y^L = f(X^L; W^{L\to L}) + f(pool(X^H, 2); W^{H\to L})$$

- $f(X ; W)$ refers to a convolution of parameters W
- average pooling operation(X, k) with kernel size k*k and stride k
- *upsample*(X, k) is a factor of k upsamplig operation via the nearest interpolation

B. Octave Convolution

Octave convolution is used as a replacement of Vanilla convolution. Octave Convolution is proposed to factorize the mixed feature maps into lower and higher frequency feature maps. The ratio of lower frequency features and lower frequency features are defined by hyper parameter α. Octconv can store and process low and high frequency feature maps. Octconv reduces spatial redundancies in CNN feature maps by down sampling spatial resolution of low frequency feature map. In Octconv low and high frequency feature maps are grouped explicitly with respect to α ratio between 0to1. Further low frequency feature maps are compressed based on scale space theory. This reduces memory utilization and computation cost.

C. Vanilla Convolution

The term "Vanilla" refers to the name given to the standard back propagation algorithm. Vanilla Back Propagation. Vanilla convolution is regular convolution, in which all input and output feature maps should be in the same spatial resolution. Octave convolution may be used as a vanilla convolution substitute. Similarly accuracy has been shown with octave convolution while saving a large number of flops needed. Model size in case of octave and vanilla convolutions is same. Vanilla convolution carries out high frequency convolution throughout all the inputs channels.

D. ReLU

ReLU is a rectified linear unit, and is a form of activation function. It is specified mathematically as $y = max(0, x)$. ReLU has been the most frequently used activation function in neural networks, in particular in CNNs. Activations of ReLU are obviously the simplest non-linear function you can use. If the input is positive then the derivative is only 1, so there is no squeezing effect from the sigmoid function that you can find on back-propagated errors. Research have shown that for large networks, ReLUs leads to more faster training. Most of the frameworks such as TensorFlow and TFLearn make it easy to use ReLUs on the hidden layers, so you don't have to implement them on your own.

E. Group-wise and Depth-wise convolutions

Octave convolution also can be used in other common variants such as depth or group-wise convolutions of the vanilla convolution. In group convolution, we simply place all four operations in groups which are appeared inside the octave convolution. Like-wise, in the depth-wise convolution, the convolution, the convolution operations are depth-wise and therefore the paths which interchange information are removed, removing two depth-wise operations. Both group Octconv and depth-wise Octconv reduces to their respective vanilla versions if the low-frequency part is not compressed.

F. Multi-scale Representation

Multi-scale representation has long been used in the extraction of local features, such as SIFT features, before the success of deep learning. Octconv can directly operates on multi-scale representation. Here input feature maps factorized into low and high frequency group. Further low frequency group compressed by 2x which contains spatially redundant information. Here intra-frequency (H->H, L->L)follows regular convolution. Regular convolution requires all output and input feature maps in spatial resolution. Inter frequency (H->L) performs pooling operation to get down-sampled result and (L->H) performs upsampling.

IV. EXPERIMENTAL RESULTS

Trained model on cifar-10 dataset which are subset of 80million tiny images. The dataset CIFAR-10 contains 60000 images of 32x32 in 10 classes with 6000 pictures in each class. Here 50,000 images are used for testing and 10,000 images are used for training. The data set consists of five training sets, each with 10000 images and one evaluation batch. Nearly 1000 pictures of each class are in the test batch.

The lots of training photos contain random images, but certain training lots that contain more pictures of one class than another. There are exactly 5,000 photos from every class in the training batches between them.
Total no.of FLOPs calculated per layer= 311040000.0
Trained on 54000 samples and validated on 6000 samples.

Table II: Table shows details of models for ResNet 50 on CIFAR-10 dataset

Model	*Total no.of parameters*	*Trainable parameters*	*Non-Trainable paameters*
Model1	34,766	33,196	1,570
Model2	34,718	33,148	1,570

Accuracy for octave convolution: 0.8995
Accuracy for normal convolution: 0.8961

V. CONCLUSION

In this paper we have suggested Octave convolution widely exists in vanilla convolution to reduce spatial redundancy in convolutional feature maps and improve the efficiency of feature maps. The lower and higher spatial frequency features used in the multi-scale representation which provides two low-resolution cost volumes. We have shown that this approach is orthogonal and complementary to existing methods, leading to a better or equivalent quality pictures that eliminate the problem collapse. Octave convolution is generic enough to substitute the standard convolution in place and can be used in major 2D and 3D CNNs without changing the design of the device. By effective communication between lower and higher frequency and by increasing their receptive field size that contributes to the acquiring of global information, Octconv improves the recognition performance.

REFERENCES

1. Han, Song, et al. "DSD: Dense-sparse-dense training for deep neural networks." *arXiv preprint arXiv:1607.04381*.
2. Luo, Jian-Hao, et al. "Thinet: pruning cnn filters for a thinner net." *IEEE transactions on pattern analysis and machine intelligence* (2018).
3. Chen, Yunpeng, et al. "Multi-fiber networks for video recognition." *Proceedings of the European Conference on Computer Vision (ECCV)*. 2018.
4. Chollet, François. "Xception: Deep learning with depthwise separable convolutions." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017.
5. VincentVanhoucke. Learning visual representations at scale. ICLR invited talk, 2014. 1, 2.
6. Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton. "Imagenet classification with deep convolutional neural networks." *Advances in neural information processing systems*. 2012.
7. Campbell, Fergus W., and John G. Robson. "Application of Fourier analysis to the visibility of gratings." *The Journal of physiology* 197.3 (1968): 551.
8. Russell L. De Valois and Karen K. De Valois. Spatial vision. Oxford psychology series, No. 14., 1988.
9. Komatsu, Ren, et al. "Octave Deep Plane-Sweeping Network: Reducing Spatial Redundancy for Learning-Based Plane-Sweeping Stereo." *IEEE Access* 7 (2019): 150306-150317.
10. Xiao, et al. "Sun3d: A database of big spaces reconstructed using sfm and object labels." *Proceedings of the IEEE International Conference on Computer Vision*. 2013.
11. Sturm, Jürgen, et al. "A benchmark for the evaluation of RGB-D SLAM systems." *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 2012.
12. Schops, Thomas, et al. "A multi-view stereo benchmark with high-resolution images and multi-camera videos." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2017.
13. Goodfellow, Ian, et al. "Generative adversarial nets." *Advances in neural information processing systems*. 2014.
14. Singh, Pravendra, et al. "Hetconv: Heterogeneous kernel-based convolutions for deep cnns." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2019.
15. Krizhevsky, Alex, and Geoffrey Hinton. *Learning multiple layers of features from tiny images*. Vol. 1. No. 4. Technical report, University of Toronto, 2009.
16. He, Kaiming, et al. "Deep residual learning for image recognition." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016.
17. Simonyan, Karen, and Andrew Zisserman. "Very deep convolutional networks for large-scale image recognition." *arXiv preprint arXiv:1409.1556* (2014).
18. Durall, Ricard, Franz-Josef Pfreundt, and Janis Keuper. "Stabilizing GANs with Octave Convolutions." *arXiv preprint arXiv:1905.12534* (2019).
19. Liu, Ziwei, et al. "Deep learning face attributes in the wild." *Proceedings of the IEEE international conference on computer vision*. 2015.
20. Pan, Zhengjun, Alistair G. Rust, and Hamid Bolouri. "Image redundancy reduction for neural network classification using discrete cosine transforms." *Proceedings of the IEEE-INNS-ENNS International Joint Conference on Neural Networks. IJCNN 2000. Neural Computing: New Challenges and Perspectives for the New Millennium*. Vol. 3. IEEE, 2000.
21. Liu, Baoyuan, et al. "Sparse convolutional neural networks." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2015.
22. Deng, Jia, et al. "Imagenet: A large-scale hierarchical image database." *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009.
23. Chen, Chun-Fu, et al. "Big-little net: An efficient multi-scale feature representation for visual and speech recognition." arXiv preprint arXiv:1807.03848 (2018).
24. Lin, Shaohui, et al. "ESPACE: Accelerating convolutional neural networks via eliminating spatial and channel redundancy." *Thirty-First AAAI Conference on Artificial Intelligence*. 2017.
25. Tran, Dat Thanh, Alexandros Iosifidis, and Moncef Gabbouj. "Improving efficiency in convolutional neural networks with multilinear filters." *Neural Networks* 105 (2018): 328-339.
26. Krizhevsky, Alex, and Geoffrey Hinton. *Learning multiple layers of features from tiny images*. Vol. 1. No. 4. Technical report, University of Toronto, 2009.
27. Netzer, Yuval, et al. "Reading digits in natural images with unsupervised feature learning." (2011).
28. Deng, Jia, et al. "Imagenet: A large-scale hierarchical image database." *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009.
29. RoyChowdhury, Aruni, et al. "Reducing duplicate filters in deep neural networks." *NIPS workshop on Deep Learning: Bridging Theory and Practice*. Vol. 1. 2017.
30. Luo, Jian-Hao, et al. "Thinet: pruning cnn filters for a thinner net." *IEEE transactions on pattern analysis and machine intelligence* (2018).
31. Shen, Yongming, Michael Ferdman, and Peter Milder. "Maximizing CNN accelerator efficiency through resource partitioning." *2017 ACM/IEEE 44th Annual International Symposium on Computer Architecture (ISCA)*. IEEE, 2017.
32. Figurnov, Mikhail, et al. "Perforatedcnns: Acceleration through elimination of redundant convolutions." *Advances in Neural Information Processing Systems*. 2016.
33. Lin, Min, Qiang Chen, and Shuicheng Yan. "Network in network." *arXiv preprint arXiv:1312.4400* (2013).
34. Chakraborty, Sinjan, et al. "Feature map reduction in CNN for handwritten digit recognition." *Recent Developments in Machine Learning and Data Analytics*. Springer, Singapore, 2019. 143-148.

AUTHORS PROFILE

A.V. Sriharsha is PhD in Privacy Preserving Data Mining and Software Architectures. He has been working as a teacher for technical degrees since 2002. His areas of interest are software architecture, data privacy, knowledge based systems, machine learning and also deep learning.

K. Yochana is PG scholar in Computer Science from the department of computer science and Engineering from Sree Vidyanikethan Engineering College Tirupati. She completed her B.tech degree from Sri Padmavathi Mahila visvavidyalayam, Tirupati, Andhra Pradesh. Her main areas are Pattern recognition and convolutional Neural Networks.