CNNs with Compact Activation Function

Jindong Wang[1], Jinchao Xu[2], and Jianqing Zhu[3]

[1] School of Mathematical Sciences, Peking University, Beijing 100871, China
jdwang@pku.edu.cn
[2] Department of Mathematics, Pennsylvania State University, University Park, PA 16802, USA
xu@math.psu.edu
http://www.personal.psu.edu/jxx1/
[3] Faculty of Science, Beijing University of Technology, Beijing 100124, China
jqzhu@emails.bjut.edu.cn

Abstract. Activation function plays an important role in neural networks. We propose to use hat activation function, namely the first order B-spline, as activation function for CNNs including MgNet and ResNet. Different from commonly used activation functions like ReLU, the hat function has a compact support and no obvious spectral bias. Although spectral bias is thought to be beneficial for generalization, we show that MgNet and ResNet with hat function still exhibit a slightly better generalization performance than CNNs with ReLU function by our experiments of classification on MNIST, CIFAR10/100 and ImageNet datasets. This indicates that CNNs without spectral bias can have a good generalization capability. We also illustrate that although hat function has a small activation area which is more likely to induce vanishing gradient problem, hat CNNs with various initialization methods still works well.

Keywords: Spectral bias · Convolutional neural network · Activation function.

1 Introduction

Activation function is an important part of neural networks. The most popular activation function is the Rectified Linear Unit (ReLU) [14]. The ReLU activation function can speed up the learning process with less computational complexity as observed in [3,12,11]. Although many other activation functions have been proposed in the last few years [13,6,9,21,17], ReLU is still the most commonly used activation function for CNNs in image classification due to its simplicity and the fact that other activation functions such as ELU [2] and GELU [9] have no significant advantage over ReLU.

Despite being heavily over-parameterized, deep neural networks have been shown to be remarkably good at generalizing to natural data. There is a phenomenon known as the spectral bias [16] or frequency principle [24,23,1] which claims that activation functions such as ReLU make the networks prioritize learning the low frequency modes and the lower frequency components of trained

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08754-7_40

networks are more robust to random parameter perturbations. This has raised the important problem of understanding the implicit regularization effect of deep neural networks and is one main reason for good generalization accuracy [15,16,20,24]. Recently, a theoretical explanation for the spectral bias of ReLU neural networks is provided in [10] by leveraging connections with the theory of finite element method and hat activation function for neural networks that different frequency components of error for hat neural networks decay at roughly the same rate and so hat function does not have the same spectral bias that has been observed for networks based on ReLU, Tanh and other activation functions.

In this paper, we consider CNN with the hat activation function which is defined as follows

$$\text{Hat(x)} = \begin{cases} x, & x \in [0,1], \\ 2-x, & x \in [1,2], \\ 0, & \text{otherwise}, \end{cases} \tag{1}$$

and is actually the B-Spline of first order. Different from ReLU function, hat function has a compact set. We use the hat activation function for CNNs including MgNet [4] and ResNet [7].

MgNet [4] is strongly connected to multigrid method and a systematic numerical study on MgNet in [5] shows its success in image classification problems and its advantages over established networks. We use MgNet in the experiment since it relates to multiscale structure that can handle the frequency variation in different resolution data. Note that hat function has a compact support which is more likely to result in the vanishing gradient problem and no spectral bias, it still obtains comparable generalization accuracy and can even perform slightly better than CNNs with ReLU activation function on MNIST, CIFAR10/100 and ImageNet datasets. This also questions whether the spectral bias is truly significant for regularization. We illustrate that the scale of hat activation function in different resolution layer is of importance and should be set properly for MgNet to adapt the frequency variation in the network. Furthermore, considering the performance of a neural network also depends on how its parameters are initialized, we also test several initialization methods for neural networks including Xavier initialization [3] and Kaiming initialization [6], the results show that all these initialization methods work well for these CNNs with hat activation function.

2 Hat function for MgNet

Different from ReLU function, the hat function has a compact support of $[0,2]$. Neural networks with hat function also have the universal approximation property [18,19,22]. Hat function is closely related to finite element method and we can adjust the compact support of this function to change its frequency. Thus, we define the following scaled hat activation function with parameter M such that

$$\text{Hat}(x;M) = \begin{cases} x, & x \in [0, \frac{M}{2}], \\ M-x, & x \in [\frac{M}{2}, M], \\ 0, & \text{otherwise}. \end{cases} \tag{2}$$

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08754-7_40

It has the advantage that its frequency can vary by changing the parameter M. It is shown in [10] that different frequency components of error for hat neural networks decay at roughly the same rate and thus hat function does not have the same spectral bias that has been observed for ReLU networks. To make full good use of this property, we introduce to use MgNet with hat activation function.

MgNet [4] is a convolutional neural network that strongly connects to multigrid methods and also has a good performance in comparison with existing CNN models [5]. The network consists of several iterative blocks shown in Figure 1 in both data space and feature space.

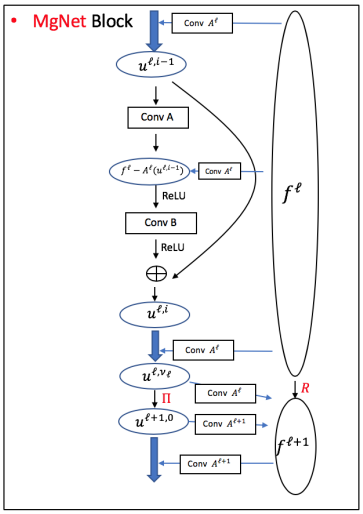

Fig. 1. MgNet Iterative block.

This block is related to the following residual correction step in multigrid method

$$u^i = u^{i-1} + B^i * (f - A^i * u^{i-1}). \tag{3}$$

where B^i, A^i are the convolution operators, f and u denote the source term (image) and solution (feature) respectively. Besides, downscaling the primary image f into coarse resolution requires the following iterative step that projects the high resolution data to the low resolution data

$$u^{\ell+1} = \Pi_\ell^{\ell+1} *_2 u^\ell, \tag{4}$$

$$f^{\ell+1} = R_\ell^{\ell+1} *_2 (f^\ell - A^\ell * u^\ell) + A^{\ell+1} * u^{\ell+1}, \tag{5}$$

where $\Pi_\ell^{\ell+1}, R_\ell^{\ell+1} *_2$ represent the convolution with stride 2. MgNet imposes some nonlinear activation function in the iterative steps above to exact the feature from image as shown in the Algorithm 1.

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08754-7_40

Algorithm 1 $Feature = \mathrm{MgNet}(f; J, \nu_1, \cdots, \nu_J)$

Initialization: $f^1 = \underline{\sigma} \circ \underline{\theta}(f)$, $u^{1,0} = 0$
for $\ell = 1 : J$ **do**
  for $i = 1 : \nu_\ell$ **do**

$$u^{\ell,i} = u^{\ell,i-1} + \underline{\sigma} \circ B^{\ell,i} * \underline{\sigma}(f^\ell - A^\ell * u^{\ell,i-1}). \tag{6}$$

  end for
  Note $u^\ell = u^{\ell,\nu_\ell}$

$$u^{\ell+1,0} = \underline{\sigma} \circ \mathit{\Pi}_\ell^{\ell+1} *_2 u^\ell \tag{7}$$

$$f^{\ell+1} = \underline{\sigma} \circ R_\ell^{\ell+1} *_2 (f^\ell - A^\ell * u^\ell) + A^{\ell+1} * u^{\ell+1,0}. \tag{8}$$

end for
Fully connected layer: $Feature = FC(Avg(u^J))$

In this algorithm, $B^{\ell,i}, A^\ell, \mathit{\Pi}_\ell^{\ell+1}, R_\ell^{\ell+1}$ are some convolution operators of a given kernel size (we often use size 3), θ is an encoding layer that increases the number of channel, Avg is the average pooling layer and σ is the activation function. The hyperparameters J and ν_i are given in advance.

We note that if we remove the variables with an underline, namely $\underline{\sigma}$ and $\underline{\theta}$ in Algorithm 1, we get exactly one classic multigrid method. From the convergence theory of multigrid method, we know that the iterative step (6) is associated with the elimination of high frequency error and with the layer getting deeper, the frequency of data gets lower. Then we can set hat activation functions with various M in different layers of the neural network to adapt this frequency variation in the network.

The MgNet model algorithm is very basic and it can be generalized in many different ways. It can also be used as a guidance to modify and extend many existing CNN models [5]. The following result shows Algorithm 1, admits the following identities

$$r^{\ell,i} = r^{\ell,i-1} - A^\ell \circ \sigma \circ B^{\ell,i} \circ \sigma(r^{\ell,i-1}), \quad i = 1 : \nu_\ell \tag{9}$$

where

$$r^{\ell,i} = f^l - A^\ell * u^{\ell,i}. \tag{10}$$

and (9) represents pre-act ResNet [8].

3 Experiments

Since MgNet has strongly connection with ResNet [7], we evaluate the performance of hat activation function on image classification for MgNet and ResNet compared with ReLU neural networks. For MgNet, we consider $J = 4$ and $\nu_1 = \nu_2 = \nu_3 = \nu_4 = 2$ as stated in Algorithm 1, thus there are four different resolution layers.

The following benchmark datasets are used: (i) MNIST, (ii) CIFAR10, (iii) CIFAR100, and (iv) ImageNet.

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08754-7_40

We consider using SGD method with the batchsize of 256 for 150 epochs. The initialization method of parameters is Kaiming's uniform intialization [6]. The initial learning rate is 0.1 with a decay rate of 0.1 per 50 epochs. The following results are the average of 3 runs. In the Table 1, the numbers [5,10,15,20] denote the scaling of hat activation in different resolution layers since the size of data have four different resolution levels in MgNet and ResNet18.

Table 1 shows that hat activation function has slightly better generalization capability than ReLU activation function for both MgNet and ResNet. To illustrate the argument of MgNet, we evaluate the performance of MgNet with different scale settings of hat activation function. As is shown in Table 2, it is better to use the hat function with larger support in the coarser resolution data which is consistent of the frequency variation of MgNet.

To exclude the influence of training process, we train the CNNs with more epochs of 300. As is shown in Table 3, the test accuracy increases both for hat MgNet and ReLU MgNet, and hat activation function still maintains slightly better generalization capability than ReLU activation function which indicates that hat activation function is truly powerful.

Since the scale of hat function is fixed which can be a potential disadvantage, we also regard these scale numbers as parameters in the network. Table 4 gives the results of trainable scale hat MgNet on CIFAR10/100 datasets and we also record the scale numbers of the model. Though using two different settings of initial scale, the results all demonstrate that it is better to use the hat function with larger support in the coarser resolution level and the support intend to be getting small during the training . The results show that the generalization accuracy of MgNet is still competitive with a much smaller support in the first few layers without adding any neurons. We also note that it is available for a combination of hat function and ReLU function for CNNs with trainable scale hat function and we can replace the activation function of the encoding layer with ReLU function.

Kaiming's initialization has been shown to work well for ReLU networks, the experiments show that hat CNNs also work well with this initialization method. Furthermore, we also consider Xavier's uniform initialization [3] for hat CNNs on CIFAR10/100 datasets. The results in Table 5 and Figure 3 show that the initialization methods make almost no difference on test accuracy but for the CIFAR100 dataset the loss of Kaiming's initialization converges slightly fast.

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08754-7_40

Table 1. Comparison of hat CNNs and ReLU CNNs for image classification.

Dataset	Model	Activation function	Test accuracy
MNIST	MgNet	ReLU	99.66
		hat-[5,10,15,20]	**99.68**
CIFAR10	MgNet	ReLU	93.13
		hat-[5,10,15,20]	**93.23**
	ResNet	ReLU	94.64
		hat-[20,15,10,5]	**94.79**
CIFAR100	MgNet	ReLU	70.85
		hat-[5,10,15,20]	**70.96**
	ResNet	ReLU	76.21
		hat-[5,10,15,20]	**76.47**
ImageNet	MgNet	ReLU	72.36
		hat-[10,20,30,40]	**72.69**

Table 2. Comparison of different scale setting of hat function for MgNet.

Dataset	Activation function	Test accuracy
MNIST	hat-[5,10,15,20]	**99.68**
	hat-[20,15,10,5]	99.64
CIFAR10	hat-[5,10,15,20]	**93.23**
	hat-[20,15,10,5]	92.87
CIFAR100	hat-[5,10,15,20]	**70.96**
	hat-[20,15,10,5]	70.56
ImageNet	hat-[10,20,30,40]	**72.69**
	hat-[40,30,20,10]	71.87

Table 3. MgNet results of 300 epochs.

Dataset	Activation function	Test accuracy
CIFAR10	hat-[5,10,15,20]	**93.97**
	hat-[20,15,10,5]	93.80
	ReLU	93.79
CIFAR100	hat-[5,10,15,20]	**72.06**
	hat-[20,15,10,5]	71.68
	ReLU	71.73

Table 4. MgNet with hat function of trainable scale (300 epochs).

Dataset	Test accuracy	Initial Scale	Final Scale
CIFAR10	94.15	[5,10,15,20]	[1.0561,1.6974,1.7712,3.2502]
	93.99	[20,15,10,5]	[1.2643,1.7570,2.8338,3.2682]
CIFAR100	72.18	[5,10,15,20]	[1.4278,2.5441,2.7431,9.6464]
	72.24	[20,15,10,5]	[2.0015,2.4351,2.7342,9.8020]

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08754-7_40

Table 5. Comparison of different initialization methods for hat-CNNs.

Dataset	Model	Activation function	Test accuracy(Kaiming)	Test accuracy(Xavier)
CIFAR10	MgNet	hat-[20,15,10,5]	93.182	93.20
		hat-[5,10,15,20]	93.233	93.26
	ResNet	hat-[20,15,10,5]	94.786	94.77
		hat-[5,10,15,20]	94.453	94.47
CIFAR100	MgNet	hat-[20,15,10,5]	70.56	70.15
		hat-[5,10,15,20]	70.963	70.89
	ResNet	hat-[20,15,10,5]	76.186	76.15
		hat-[5,10,15,20]	75.996	76.22

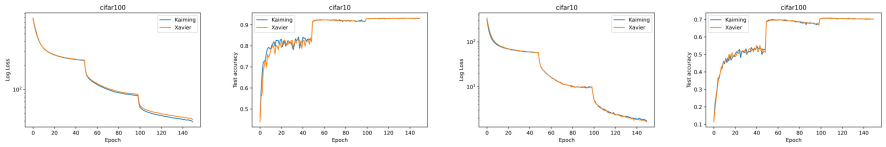

Fig. 2. Comparison of loss curve and test accuracy curve of MgNet for CIFAR10 and CIFAR100 datasets versus initialization methods.

4 Conclusion

We introduce the hat activation function, a compact function for CNNs and evaluate its performance on several datasets in this paper. The results show that hat activation function which has a compact activation area still has competitive performance in comparison with ReLU activation function for MgNet and ResNet although it does not have those properties of ReLU which are deemed to be important. Besides, activation function with a small compact set can cause gradient vanishing easily but this has no influence on the performances of CNNs with hat function. Specifically, from the experiments we note that the scale setting of hat activation function also influences the performance, which is related to the frequency variation in the network. Furthermore, commonly used initialization methods are also shown to be viable for hat CNNs. These numerous experiments show that hat function is indeed a viable choice of activation functions for CNNs and indicate the spectral bias is not significant for generalization accuracy.

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08754-7_40

Acknowledgments

The work of Jinchao Xu is supported in part by the National Science Foundation (Grant No. DMS-2111387). The work of Jianqing Zhu is supported in part by Beijing Natural Science Foundation (Grant No. Z200002). The work of Jindong Wang is supported in part by High Performance Computing Platform of Peking University.

References

1. Basri, R., Galun, M., Geifman, A., Jacobs, D., Kasten, Y., Kritchman, S.: Frequency bias in neural networks for input of non-uniform density. In: International Conference on Machine Learning. pp. 685–694. PMLR (2020)
2. Clevert, D.A., Unterthiner, T., Hochreiter, S.: Fast and accurate deep network learning by exponential linear units (elus). arXiv preprint arXiv:1511.07289 (2015)
3. Glorot, X., Bengio, Y.: Understanding the difficulty of training deep feedforward neural networks. In: Proceedings of the thirteenth international conference on artificial intelligence and statistics. pp. 249–256. JMLR Workshop and Conference Proceedings (2010)
4. He, J., Xu, J.: MgNet: A unified framework of multigrid and convolutional neural network. Science china mathematics **62**(7), 1331–1354 (2019)
5. He, J., Xu, J., Zhang, L., Zhu, J.: An interpretive constrained linear model for ResNet and MgNet. arXiv preprint arXiv:2112.07441 (2021)
6. He, K., Zhang, X., Ren, S., Sun, J.: Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In: Proceedings of the IEEE international conference on computer vision. pp. 1026–1034 (2015)
7. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
8. He, K., Zhang, X., Ren, S., Sun, J.: Identity mappings in deep residual networks. In: European conference on computer vision. pp. 630–645. Springer (2016)
9. Hendrycks, D., Gimpel, K.: Gaussian error linear units (gelus). arXiv preprint arXiv:1606.08415 (2016)
10. Hong, Q., Siegel, J., Tan, Q., Xu, J.: On the activation function dependence of the spectral bias of neural networks. preprint (2022)
11. Klambauer, G., Unterthiner, T., Mayr, A., Hochreiter, S.: Self-normalizing neural networks. Advances in neural information processing systems **30** (2017)
12. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems **25** (2012)
13. Maas, A.L., Hannun, A.Y., Ng, A.Y., et al.: Rectifier nonlinearities improve neural network acoustic models. In: Proc. icml. vol. 30, p. 3. Citeseer (2013)
14. Nair, V., Hinton, G.E.: Rectified linear units improve restricted boltzmann machines. In: ICML (2010)
15. Poggio, T., Kawaguchi, K., Liao, Q., Miranda, B., Rosasco, L., Boix, X., Hidary, J., Mhaskar, H.: Theory of deep learning iii: the non-overfitting puzzle. CBMM Memo **73** (2018)

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08754-7_40

16. Rahaman, N., Baratin, A., Arpit, D., Draxler, F., Lin, M., Hamprecht, F., Bengio, Y., Courville, A.: On the spectral bias of neural networks. In: International Conference on Machine Learning. pp. 5301–5310. PMLR (2019)
17. Ramachandran, P., Zoph, B., Le, Q.V.: Searching for activation functions. arXiv preprint arXiv:1710.05941 (2017)
18. Siegel, J.W., Xu, J.: Characterization of the variation spaces corresponding to shallow neural networks. arXiv preprint arXiv:2106.15002 (2021)
19. Siegel, J.W., Xu, J.: Improved approximation properties of dictionaries and applications to neural networks. arXiv preprint arXiv:2101.12365 (2021)
20. Soudry, D., Hoffer, E., Nacson, M.S., Gunasekar, S., Srebro, N.: The implicit bias of gradient descent on separable data. The Journal of Machine Learning Research **19**(1), 2822–2878 (2018)
21. Trottier, L., Giguere, P., Chaib-Draa, B.: Parametric exponential linear unit for deep convolutional neural networks. In: 2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA). pp. 207–214. IEEE (2017)
22. Xu, J.: The finite neuron method and convergence analysis. arXiv preprint arXiv:2010.01458 (2020)
23. Xu, Z.Q.J., Zhang, Y., Luo, T., Xiao, Y., Ma, Z.: Frequency principle: Fourier analysis sheds light on deep neural networks. arXiv preprint arXiv:1901.06523 (2019)
24. Xu, Z.J.: Understanding training and generalization in deep learning by fourier analysis. arXiv preprint arXiv:1808.04295 (2018)

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08754-7_40