A new approach for upscaling document images for improving their quality

Ram Krishna Pandey
Indian Institute Of Science
Bangalore, India
Email: ramp@iisc.ac.in

Shishira R Maiya
Ramaiah Institute of Technology
Bangalore, India
Email: shishira.maiya96@gmail.com

A G Ramakrishnan
Indian Institute Of Science
Bangalore, India
Email: agr@iisc.ac.in

***Abstract*—One of the issues faced by optical character recognition (OCR) softwares is the input document images being not of good quality. So research into the methods of enhancing the document images, before presenting them to OCR softwares, is of utmost importance. The objective is to demonstrate a method of generating a high resolution document image, given a low resolution image. We propose a new method for improving the spatial resolution of document images. Here, we have built a deep neural network based model that utilizes the traditional interpolation methods, takes the best features from them and reconstructs a high resolution image from these features. This is achieved using a convolutional neural network (CNN). The CNN learns a high resolution patch from a corresponding low resolution patch, as a weighted non-linear combination of the outputs of different interpolation techniques. We call our technique as nonlinear fusion of multiple interpolations (NFMI). The NFMI method ensures that the model learns only the best features that can be extracted from all the interpolation techniques combined together. The use of traditional interpolation methods makes sure that the NFMI technique is not computationally expensive. Results on test images show a relative improvement of 54% in word recognition accuracy by OCR over the best interpolation technique for doubling the spatial resolution and 33% for quadrupling the resolution.**

***Index Terms*—Super resolution, CNN, DNN, OCR, document images, deep learning, interpolation, feature extraction, image reconstruction, neural networks.**

I. Introduction

Super resolution refers to the class of techniques used to enhance the quality of an image, while retaining its information content. Since its conception in the 1980's, these techniques have been extensively used for the enhancement of video and natural images, but not much work has been reported on document images. The enhancement of document images is an important problem, considering the fact that the quality of these images plays a crucial role in the accuracy of optical character recognition (OCR) softwares, which take an image of a printed document and convert it into an editable document [1]- [6]. For example, if there is a task focused on converting a scanned document image to text, even the slightest variation in the quality of the input image can lead to spurious output results, reducing the effectiveness of the process. In certain digitization projects, a low scanning resolution was used to minimize bandwidth requirements in sending the document images to the volunteers [7]. So, an improvement in image quality will positively impact the accuracy of an OCR engine [8]. Our work reported in this paper mainly deals with document images [9].

Historically, interpolation techniques such as bicubic, bilinear, nearest neighbor and spline have been used for the super resolution of images. These techniques proved to be very effective on natural images. Each of these methods can obtain the high resolution version of the document image. These are straightforward methods of going from a low resolution to a high resolution manifold of document images. Motivated by the way deep neural networks work, where each layer is some kind of representation of the input data, we have used convolutional neural networks (CNN) to introduce nonlinearity to the model. Since the normal CNN can decrease or retain the dimentionality of the features, we have used interpolation techniques to obtain features in the high resolution image space and then pass these to the CNN to obtain more complex features and so on. The idea behind this is to weigh the interpolation of a pixel in high resolution space based on the quality of the document image we desire.

Bilinear interpolation uses the four nearest pixel values along the diagonals, and calculates a weighted average of the attributes of those pixels. This action averts the possibility of the effect of one pixel being higher than the other in the upscaled image. This method captures the texture information of the given input image, which is useful in the construction of its high resolution counterpart. Nearest neighbour interpolation yields a piecewise constant interpolation based on its neighbours. This method captures the local pixel information, which helps in getting details for the enhanced output image. Bicubic interpolation, unlike the other methods, takes into account a 4X4 grid or 16 pixels. The output from bicubic method is smoother than those of the other methods, due to the use of the gradient information of the pixels.

Each of these methods, due to the properties listed above, have proved to be successful in their own right. Though these methods provide a computationally less expensive way of upscaling an image, the individual results of none of these methods is satisfactory for document images. This is due to the fact that each of them is good at extracting a particular feature, but ignore other features, which might be relevant in the upscaled document image. The model proposed in this paper utilizes a nonlinearly weighted combination of the best features of these methods to construct a high resolution image.

Hence, we call our method as nonlinear fusion of multiple interpolations (NFMI).

II. Problem Definition

This research problem arose from an actual industry, where they had to deal with the recognition of a huge number of documents, which had been poorly scanned in the binary mode, and also, the original paper documents were no longer available. The problem can simply be stated as the reconstruction of a high resolution output image from a given low resolution input document image, which will lead to increased recognition accuracy by existing good OCRs. In this paper, we propose a model to upscale an input document image by a factor of two or four.

Upscaling by a factor of two: In this method, the input image is separately passed through three interpolation techniques, namely bicubic, bilinear and nearest neighbor. The outputs of these three methods are combined together and input to the CNN. The CNN model consists of 3 layers, with rectified linear unit (ReLU) layers placed in between the layers to obtain nonlinearity. The first layer of the CNN takes in the combined tensor. The filters in this layer convolve with the input tensor to produce feature maps of the input patches. These feature maps are then presented to the subsequent layers. The output from the final CNN layer, which has only one 3X3 filter, is a single upscaled (2X) patch of the corresponding high resolution output.

Upscaling by a factor of four: The task is exactly the same as the previous one; the only difference is that the input image is to be upscaled by a factor of 4 by the pre-existing interpolation techniques.

III. Related Work in the literature

Image super-resolution is a well-known, highly ill-posed problem. The upscaling of images can be performed by simple interpolation techniques like bicubic, bilinear and nearest neighbor. However, some information is lost in the reconstruction using these techniques, such as the high frequency component. Super-resolution (SR) techniques can be broadly divided into two: (a) multi frame (b) single image. In multi frame image super-resolution, multiple images are required to obtain the high resolution image. On the other hand, in single image SR, only the input image is required to construct a high resolution image. In most of the cases, we don't have multiple images with sub-pixel alignment to reconstruct the HR image.

Recently, various learning based techniques have been developed to obtain the HR image. Almost all of these methods work on natural images; their basic aims are almost the same: i.e. to learn a mapping function which can obtain the correspondence between a low resolution manifold and its corresponding high resolution counterpart. [10] proposes a convolutional neural network based natural image super-resolution, which shows significant improvement in terms of the peak signal-to-noise ratio (PSNR) with respect to the downsampled version of the original images. In this work, the downsampled version of the image is upscaled by bicubic interpolation and a three layer CNN is used to obtain the high resolution image. The main disadvantage is that it loses the high frequency components in the reconstructed image. Document images have more edges which need to be preserved and less texture information. This problem of obtaining a high resolution document image from a single low resolution image was first reported in the paper titled "Efficient Document-Image Super-Resolution using Convolutional Neural Network" [11].

IV. Contributions of this work

In this paper, we address the problem of single image super resolution in a way different from those addressed in the literature. In this paper, we have used a convolutional neural network (CNN) based architecture. Usually in CNN, the size of the output decreases or remains the same. The challenge is how one can use CNNs to upscale the image. So, we have used traditional techniques of [12], [13] and [14] and combine the outputs of these techniques using a CNN to learn a mapping function to obtain a high-resolution image. Our proposed technique takes advantage of the existing interpolation techniques. This task can be accomplished by using one or more layers of transposed convolution [15] or sub-pixel convolution [16]. It means that we have obtained each pixel of the high resolution image as a nonlinear weighted combination of the outputs of the existing interpolation methods. This technique gives performance superior to the interpolation techniques applied individually, as can be seen from the results reported later.

V. Dataset created for the study

Since we are addressing the problem of Tamil document image enhancement, and there is no publicly available dataset for our task, we have created our own dataset. The dataset contains a total of 2,015,815 binary image patch pairs.

A. For upscaling by factor of 2

Training images used to create high resolution patches are scanned at 200 dots per inch. For obtaining the corresponding low resolution patches, alternate pixels are chosen from the above images.

B. For upscaling by a factor of 4

Images used to create high resolution training patches are scanned at 300 dots per inch and the corresponding low resolution patches are obtained by choosing one pixel after every 3 pixels.

VI. Details of implementation

The details of the architecture proposed are given in Figs. 1 and 2. The architecture was implemented in python's keras [17] library with tensorflow [18] backend.

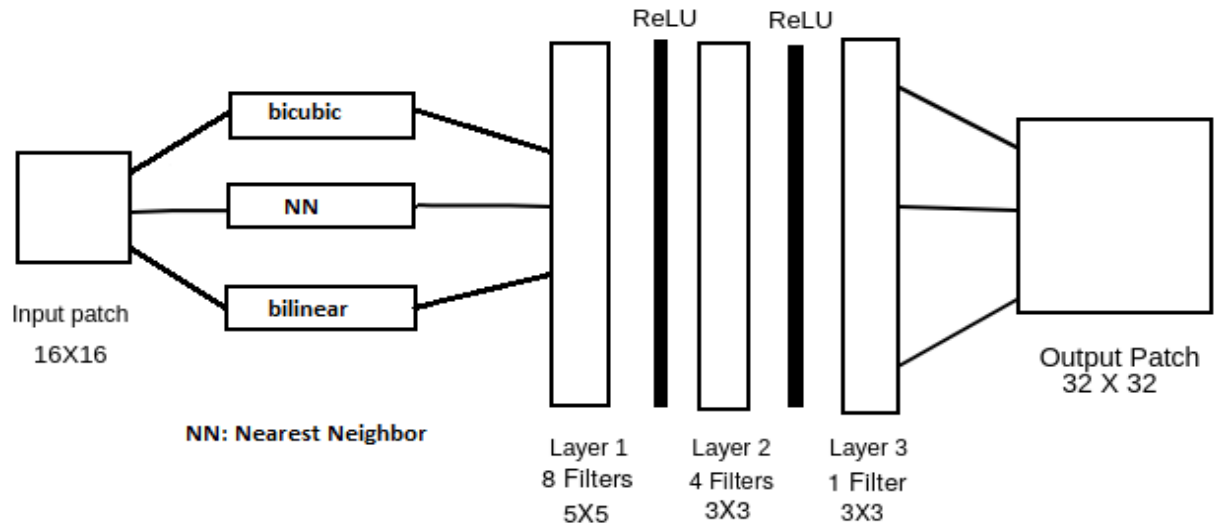


Fig. 1. Architecture proposed for nonlinear fusion of multiple interpolations for upscaling a low resolution document image by a factor of 2, increasing its recognition potential. Spatial resolutions of the input and output images used for training the network are 100 and 200 dpi, respectively.

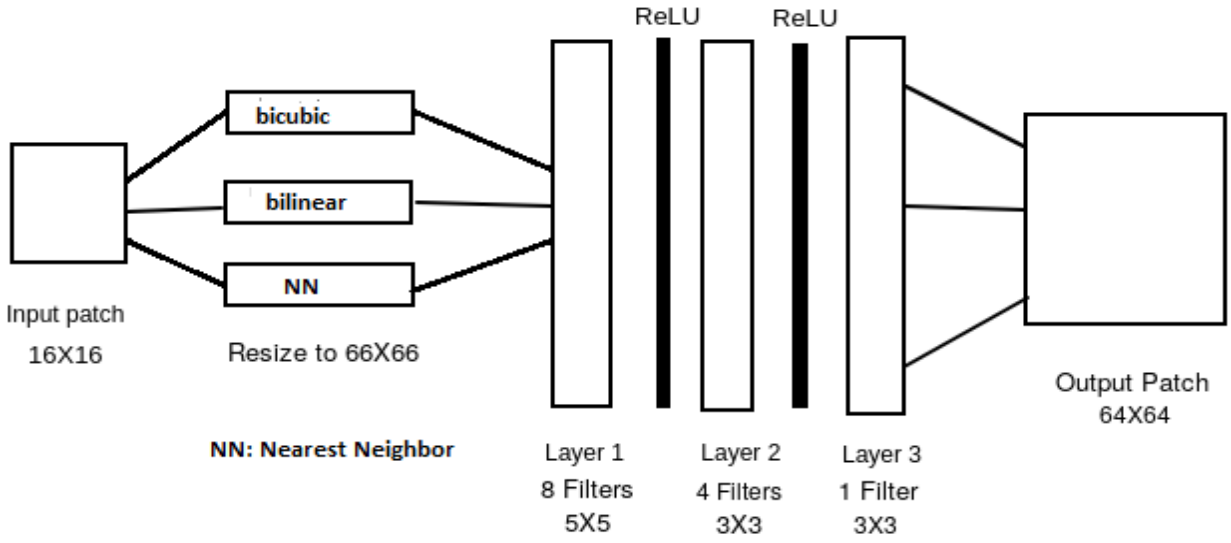


Fig. 2. Architecture proposed for nonlinear fusion of multiple interpolations for upscaling a low resolution document image by a factor of 4, while increasing its recognition potential. Resolutions of the input and output images used for training the network are 75 and 300 dpi, respectively.

A. Upscaling by a factor 2

The input image is first upscaled by 2 times independently by the application of the three interpolation techniques, namely, bicubic, bilinear and nearest neighbor. The three outputs are passed on as features to the CNN. The CNN architecture has 8 filters of size 5×5 in the first layer and 4 filters of size 3×3 in the second layer. The final layer has a single filter of size 3×3, since we need only one output at the output layer.

B. Upscaling by a factor of 4

The input image is upscaled by a factor of 4 employing the existing interpolation techniques, and the three outputs are passed on as features to the CNN. The CNN architecture after upscaling is the same as that of the previous one used for upscaling by a factor of 2.

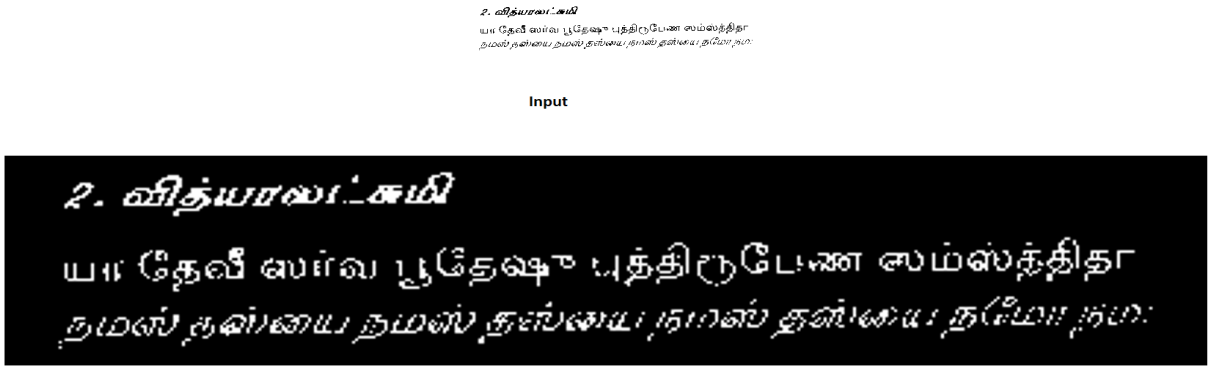
C. Details of training

For training the network, stochastic gradient descent with momentum and normal back propagation is used [19], [20], [21]. The parameters used during training are: batch size = 32, learning rate = 0.02, momentum = 0.88.

VII. Results

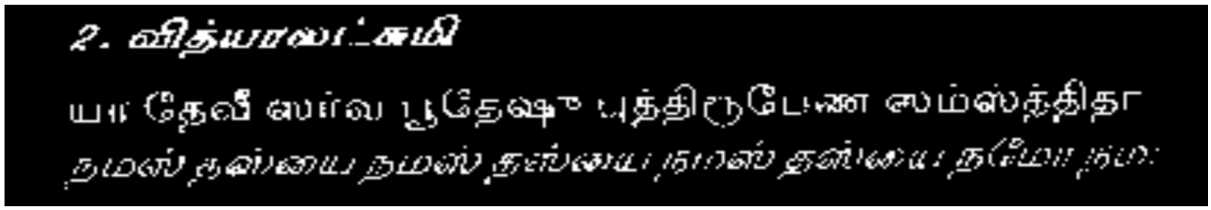
Figures 3 and 4 show the input image to the model which is scanned at 100 dot per inch (dpi) and the results obtained by interpolating with the various available techniques i.e bicubic, bilinear and nearest neighbor and our reconstruction which is upscaled by a factor of 2 and 4, respectively. Since Gamma

Fig. 3. Input is a sample Tamil image scanned at 100 dpi. The image output by the proposed NFMI method is compared with the output images of bicubic, bilinear and nearest neighbor interpolations, all for an upscale factor of 2.

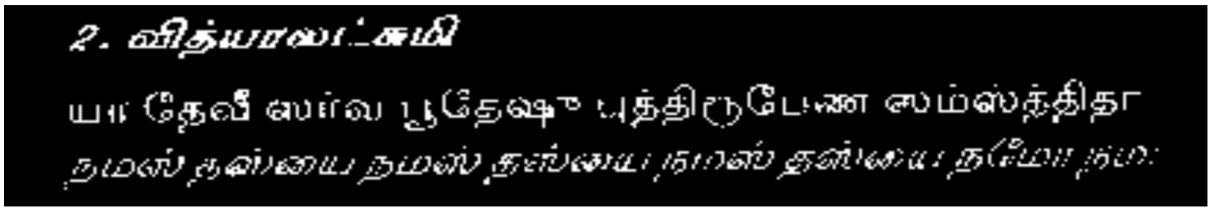


Input

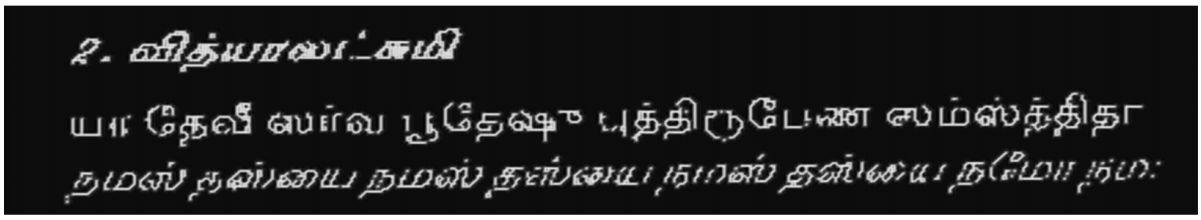
Bicubic



Bilinear



Nearest Neighbor



Our Method

Fig. 4. Input is a sample Tamil image scanned at 100 dpi. The image output by the proposed NFMI method is compared with the output images of bicubic, bilinear and nearest neighbor interpolations, all for an upscale factor of 4.

TABLE I

WORD LEVEL ACCURACY IN % OBTAINED BY THE OCR ON THE INPUT IMAGES (100 DPI) AND ON THE IMAGES OUTPUT BY THE DIFFERENT INTERPOLATION TECHNIQUES AND OUR METHOD OF NONLINEAR FUSION OF MULTIPLE INTERPOLATIONS FOR AN UPSCALE FACTOR OF 2.

Input	Bicubic-$\times 2$	Bilinear-$\times 2$	Nearest neighbor-$\times 2$	NFMI reconstruction-$\times 2$	NFMIγ-Reconstructed-$\times 2$
4.09	13.13	10.47	13.1	14.29	20.20

TABLE II

WORD LEVEL ACCURACY IN % OBTAINED BY THE OCR ON THE INPUT IMAGES (100 DPI) AND ON THE IMAGES OUTPUT BY THE DIFFERENT INTERPOLATION TECHNIQUES AND OUR METHOD OF NONLINEAR FUSION OF MULTIPLE INTERPOLATIONS FOR AN UPSCALE FACTOR OF 4.

Input	Bicubic-$\times 4$	Bilinear-$\times 4$	Nearest neighbor-$\times 4$	NFMI reconstruction-$\times 4$	NFMIγ-Reconstructed-$\times 4$
4.09	15.38	7.69	16.48	21.98	21.98

correction cannot be applied on binary images, it is performed only on the output images to increase the recognition potential of the gray images obtained from the model. The OCR accuracy is maximum for a particular value of gamma, the details of which can be found in [22]. Tables I and II compare the word level accuracies obtained by the Tamil OCR on the outputs of our method with those of the interpolation techniques, for upscaling factors of 2 and 4, respectively.

In upscaling by a factor of 4, γ-correction is not able to obtain better results than the output reconstructed.

VIII. CONCLUSION

We have proposed a novel DNN-based CNN architecture, which nonlinearly combines the outputs of multiple interpolation techniques. It obtains better results in terms of OCR word level accuracy and the output images visually look better than those of the individual interpolation techniques. We have tried our architecture only for Tamil document images due to the unavailability of standard datasets. Our model can be applied for upscaling document images in any language (script). It can also be applied to natural images, if trained on natural images. Further, we will try to explore multiple ways of creating the dataset so that the model generalizes well and can be tuned to perform better by selecting different number of filters and hidden layers. Our model can scale to other languages and resolutions by mixing the data from different languages and resolutions.

IX. ACKNOWLEDGMENT

The first author thanks Mr. Madhavaraj A for all the fruitful discussions he had regarding this work. The authors thank Mr. A. Elangovan, Managing Director, Cadgraf Digitals Private Ltd., Chennai for challenging us with this real-life industrial problem to work with.

REFERENCES

[1] K. G. Aparna and A. G. Ramakrishnan, "A complete Tamil Optical Character Recognition System," Proc. 5th IAPR Workshop on Document Analysis Systems DAS-02, Princeton, NJ, August 19-21, 2002, pp. 53-57.

[2] Shiva Kumar H R and A G Ramakrishnan, A tool that converted 200 Tamil books for use by blind students, Proc. 12-th International Tamil Internet Conf., Kuala Lumpur, Malaysia, Aug. 15-18, 2013.

[3] Manthan award 2014 for the project, Gift of new abilities. http://manthanaward.org/e-inclusion-accessibilty-winner-2014/

[4] Vijay Kumar B and A G Ramakrishnan, Radial basis function and subspace approach for printed Kannada text recognition, Proc. IEEE ICASSP-04, May 17-21, 2004, Montreal, Canada, Vol 5, pp. 321-324.

[5] Vijay Kumar and A.G. Ramakrishnan, "Machine recognition of printed Kannada text," Proc. 5th IAPR Workshop on Document Analysis Systems (DAS-02), Aug 19-21, 2002, Springer Verlag, Berlin, pp. 37-48.

[6] R S Umesh, Peeta Basa Pati and A G Ramakrishnan, Design of a bilingual Kannada-English OCR, in the book Guide to OCR for Indic Scripts: Document Recognition and Retrieval Springer, 2009 in the Advances in Pattern Recognition Series. Ed: Venu Govindaraju and Setlur Srirangaraj. pp. 97-124. ISBN: 978-1-84800-330-9

[7] Project Madurai for ancient Tamil literary works. http://www.projectmadurai.org/

[8] A Madhavaraj, A G Ramakrishnan, H R Shiva Kumar, Nagaraj Bhat, "Improved recognition of aged Kannada documents by effective segmentation of merged characters", Proc. Tenth Int. Conf. on Signal Process. Commun. (SPCOM), Bangalore, July 22-24, 2014.

[9] Ram Krishna Pandey, and A. G. Ramakrishnan. "Language Independent Single Document Image Super-Resolution using CNN for improved recognition." arXiv preprint arXiv:1701.08835 (2017).

[10] Dong, Chao, Chen Change Loy, Kaiming He, and Xiaoou Tang. "Image super-resolution using deep convolutional networks." IEEE Trans. Pattern Analysis and Machine Intelligence, 38, no. 2 (2016): 295-307.

[11] Ram Krishna Pandey, and A. G. Ramakrishnan. "Efficient Document-Image Super-Resolution Using Convolutional Neural Network" in press, Sadhana, 2017.

[12] Keys R, "Cubic convolution interpolation for digital image processing". IEEE Trans. Acoustics, Speech, Sig. Process., 1981: 29(6), 1153-1160.

[13] Han, Dianyuan. "Comparison of commonly used image interpolation methods." Proc. 2nd International Conf. Computer Science and Electronics Engineerings (ICCSEE), pp. 1556-1559. 2013.

[14] Yang, C. S., Kao, S. P., Lee, F. B., and Hung, P. S. (2004, July). "Twelve different interpolation methods: A case study of Surfer 8.0". Proc. XXth ISPRS Congress (Vol. 35, pp. 778-785.

[15] Xu, Li, Jimmy SJ Ren, Ce Liu, and Jiaya Jia. "Deep convolutional neural network for image deconvolution." In Advances in Neural Information Processing Systems, pp. 1790-1798. 2014.

[16] Shi, Wenzhe, Jose Caballero, Ferenc Huszr, Johannes Totz, Andrew P. Aitken, Rob Bishop, Daniel Rueckert, and Zehan Wang. "Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network." Proc. IEEE Conf. Computer Vision and Pattern Recognition, pp. 1874-1883. 2016.

[17] Chollet, Franois. "Keras." (2015).

[18] Abadi, Martn, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado et al. "Tensorflow: Large-scale machine learning on heterogeneous distributed systems." arXiv preprint arXiv:1603.04467 (2016).

[19] LeCun, Yann, et al. "Backpropagation applied to handwritten zip code recognition." Neural Computation 1.4 (1989): 541-551.

[20] Bottou, L. (2010). "Large-scale machine learning with stochastic gradient descent". Proc. COMPSTAT'2010 (pp. 177-186). Physica-Verlag HD.

[21] Goh, Gabriel. "Why Momentum Really Works." Distill 2.4 (2017): e6.

[22] Kumar, Deepak, and A. G. Ramakrishnan. "Power-law transformation for enhanced recognition of born-digital word images." Proc. Intern. Conf. Signal Process. Commun. (SPCOM), pp. 1-5. IEEE, 2012.