中國機械工程學刊第四十二卷第一期第 23~31 頁(民國一百一十年)
Journal of the Chinese Society of Mechanical Engineers, Vol.42, No.1, pp 23~31 (2021)

Image Recognition based on Dynamic Highway Networks

Sui-Hsien Wang*, Wei-Zhi Lin*and Han-Pang Huang*

Keywords：dynamic highway networks, machine learning, highway networks.

ABSTRACT

With the development of machine learning technology, more and more complex networks are developed. For those networks, determining the hyperparameters is important so that they can provide the best performance under the structure of neural network. However, more parameters should be decided in complex networks. This paper is focused on developing a structure of neural networks which can tune the width in each layer based on the utility of neurons automatically. In order to realize this function, a new structure of neural network called convolution neural network based dynamic highway network is proposed to deal with image recognition problem. With the self-adjusting method, near optimal structure and few parameters are required for training to achieve the same and even better performance which uses more neurons.

INTRODUCTION

With the development of machine learning, people have tried to use it to resolve increasingly difficult problems. More and more complex networks were designed and more neurons and layers were used (Guo et al., 2016). Suffering from the vanishing gradient problem, different structures, and activation functions were developed. Residual learning (He, Zhang, Ren, and Sun, 2016; Wu, Zhong, and Liu, 2017), SkipNet (Wang, Yu, Dou, and Gonzalez, 2017) and highway networks (Srivastava, Greff, and Schmidhuber, 2015) were tried out to bypass some of the layers and transmit information to deeper layers. Using this approach, the algorithm can avoid the vanishing gradient and construct a deeper network to address more difficult problems.

However, the complicated structure, enormous was required for training. Dynamic neural networks were proposed and these provided two ways to overcome this problem. First, the parameters in major networks were generated by other minor networks so that the number of parameters used can be reduced. Second, the structure, including the number of neurons and the connections with each other, was modified by specific rules, making it possible to use fewer parameters to achieve the same performance of heavy networks.

In this research, a new convolution neural network (CNN) based dynamic neural network that uses the basic idea of highway networks and residual networks is proposed. It offers an efficient method to generate new neurons and grow from a small network. As a result, it needs less time for training and uses fewer parameters to provide high performance for image recognition applications.

The remainder of the paper is organized as follows. Section II reviews the state-of-the-art highway network application. Section III presents the CNN-based dynamic highway networks. Section IV conducts simulations and experiments. MNIST and CIFAR10 datasets are used to test the proposed CNN-based dynamic highway networks. Section V summarize the works of this paper and suggest future works.

RELATED WORKS

CNN has achieved great success in the field of computer vision. In each ImageNet Large Scale Visual Recognition Competition (ILSVRC), there were many structures such as AlexNet (Krizhevsky, Sutskever, and Hinton, 2012), GoogleNet (Szegedy et al., 2015) and ResNet (He et al., 2016; Wu et al., 2017) which had improved the performance. In the deep learning structure, the number of depth layer was increased to solve the complex recognition tasks. The vanishing gradient makes the training process more difficult (Bengio, Simard, and Frasconi, 1994). Highway network (Srivastava et al., 2015) is one of the methods to solve this problem by using bypassing the information from input to the output layer. It was successfully used in computer vision such as image classification (Oyedotun, Shabayek, Aouada, and Ottersten, 2018), synthetic aperture radar (SAR) target classification (Lin, Ji, Kang, Leng, and Zou,

Paper Received October, 2019. Revised May, 2020. Accepted June, 2020. Author for Correspondence:Han-Pang Huang.

* *Student, Department of Mechanical Engineering, National Taiwan University, No. 1, Sec. 4, Roosevelt Rd., Taipei 10617, TAIWAN (R.O.C.).*

** *Professor, Department of Mechanical Engineering, National Taiwan University, No. 1, Sec. 4, Roosevelt Rd., Taipei 10617, TAIWAN (R.O.C.).*

2017) and single image super-resolution (SISR) (Li, Bare, Yan, Feng, and Yao, 2018). Lin et al. (Lin et al., 2017) used the convolutional highway unit to train a synthetic aperture radar (SAR) target classification system with the limited SAR data. And in SISR, Li et al. (Li et al., 2018) proposed the Highway Networks Super Resolution (HNSR) to reconstruct the high-resolution image with a part of loss function called structural similarity index (SSIM). In (Oyedotun et al., 2018), the authors added gate constraints to reformulate the highway blocks that learned feature transformation as model training progresses. Although those models outperformed the original highway units, those models had no discussion about the effects of the number of neurons in their proposed structures.

Zagoruyko et al (Zagoruyko and Komodakis, 2016) analyzed the effects of width for the model. They proposed a novel structure called wide residual networks (WRNs) which decrease the number of layers and increase the number of neurons for residual networks. They showed that the performance of WRN with different number of layers was better than more deep networks. However, the WRN used parameters as many as the deep neural network. It requires a lot of memory space to optimize and store the parameters. To reduce the number of the parameters, Lu and Renals (Lu and Renals, 2017) trained a small-footprint highway network to achieve better recognition accuracy with much less model parameters than the classic deep neural network. Nevertheless, there is no suggestion to choose how many parameters can be reduced to suit a particular task.

However, those applications still used the trial and error to choose the width of each layer. It takes a lot of time to find the optimal number of parameters. In this paper, we propose a new structure called CNN-based dynamic highway network. The structure with a self-adjusting gate can adjust the width of each layer that has been optimized.

CNN BASED DYNAMIC HIGHWAY NETWORKS

In this section, a new structure of neural networks called CNN based dynamic highway networks is proposed. It offers several advantages, such as reducing the time required for training and the memory occupied.

Highway Networks

Srivastava, Greff, and Schmidhuber (Szegedy et al., 2015) proposed highway networks which involved the idea of bypassing some layers and transmitting information to the deeper layers. In traditional shallow neural networks with L layers, the transform in each layer can be formulated as follows.

$$y = H(\mathrm{W}_H \mathrm{x} + b_H) \tag{1}$$

where $\mathrm{x} \in \mathrm{R}^m, y \in \mathrm{R}^n$ denote the input and output of the layer respectively, $\mathrm{W}_H \in \mathrm{R}^{n \times m}, b_H \in \mathrm{R}^n$ denote the weight and bias in the layer and H is the activation function which is usually a sigmoid function.

In highway networks, an additional transform $T(\mathrm{W}_T \mathrm{x} + b_T)$ is used to control the weights of the combination of input x and the output of transform $H(\mathrm{W}_H \mathrm{x} + b_H)$. The detailed function is shown in equation (2) and illustrated in Figure 1.

$$y = H(\mathrm{W}_H \mathrm{x} + b_H) \cdot T(\mathrm{W}_T \mathrm{x} + b_T) + \mathrm{x} \cdot (1 - T(\mathrm{W}_T \mathrm{x} + b_T)), \tag{2}$$

where x, y, W_H, b_H follow the definition above, $\mathrm{W}_T \in \mathrm{R}^{n \times m}, b_T \in \mathrm{R}^n$ denote the weights and bias in the transform $T(\mathrm{W}_T \mathrm{x} + b_T)$, H is an activation function which is a rectified linear unit here, and T is an activation function which is a sigmoid function here, the dot operator $(\cdot)$ denotes element-wise multiplication, 1 denotes the vector of one. The detailed functions of T and H are shown in equation (3) and equation (4).

$$T(\mathrm{x}) = \frac{1}{1 + e^{-x}}, \tag{3}$$

$$H(\mathrm{x}) = \max(0, \mathrm{x}), \tag{4}$$

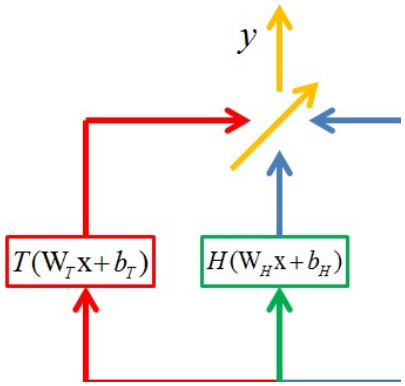

Fig. 1 Illustration of highway networks

In addition, equation (2) shows that y is a linear combination of input x and the output of transform $H(\mathrm{W}_H \mathrm{x} + b_H)$ because the output of function T is between 0 and 1.

Highway networks have the advantage that they can train very deep networks by backpropagation without two-stage training. During training, backpropagation gives the direction to reduce loss. The function of backpropagation can be derived as equation (5) and equation (6).

$$\frac{\partial L}{\partial W_{H_{i,j}}} = \frac{\partial L}{\partial y_i} \frac{\partial y_i}{\partial H_i} \frac{\partial H_i}{\partial W_{H_{i,j}}}, \tag{5}$$

$$\frac{\partial L}{\partial W_{T_{i,j}}} = \frac{\partial L}{\partial y_i}\frac{\partial y_i}{\partial T_i}\frac{\partial T_i}{\partial W_{H_{i,j}}}, \tag{6}$$

where L is the loss and the subscripts i and j denote the j^{th} variable in the i^{th} layer. Equation (5) shows that $W_{H_{i,j}}$ changes as plain feedforward networks do. On the other hand, equation (6) displays how $W_{T_{i,j}}$ changes to reduce loss. To go back to the idea of the output of $T(\mathrm{W}_T\mathrm{x}+b_T)$, it is the weights of the combination of input x and the output of transform $H(\mathrm{W}_H\mathrm{x}+b_H)$, and equation (6) tunes the weights to reduce loss. As a result, we can consider the weights to be tuned according to the utility of input x and the output of transform $H(\mathrm{W}_H\mathrm{x}+b_H)$ by backpropagation. Although the outputs of $H(\mathrm{W}_H\mathrm{x}+b_H)$ and $T(\mathrm{W}_T\mathrm{x}+b_T)$ change with epochs at the same time, backpropagation provides the best direction to reduce loss. In other words, the output of $T(\mathrm{W}_T\mathrm{x}+b_T)$ tends to be a small value if the weights in $H(\mathrm{W}_H\mathrm{x}+b_H)$ still result in huge loss after backpropagation.

Dynamic Highway Networks

Based on the idea of the output of $T(\mathrm{W}_T\mathrm{x}+b_T)$, the utility of input x and the output of transform $H(\mathrm{W}_H\mathrm{x}+b_H)$ can be compared. Therefore, a new structure of highway networks is designed so that it brings some merits. The structure is shown in equation (7), equation (8) and Figure 2.

$$d = T_1(\mathrm{W}_{T_1}\mathrm{x}+b_{T_1})\cdot H_1(\mathrm{W}_{H_1}\mathrm{x}+\mathrm{b}_{H_1}), \tag{7}$$

$$y = T_2(\mathrm{W}_{T_2}\mathrm{x}+b_{T_2})\cdot H_2(\mathrm{W}_{H_2}\,d+b_{H_2}) + (1-T_2(\mathrm{W}_{T_2}\mathrm{x}+b_{T_2}))\cdot \mathrm{x}, \tag{8}$$

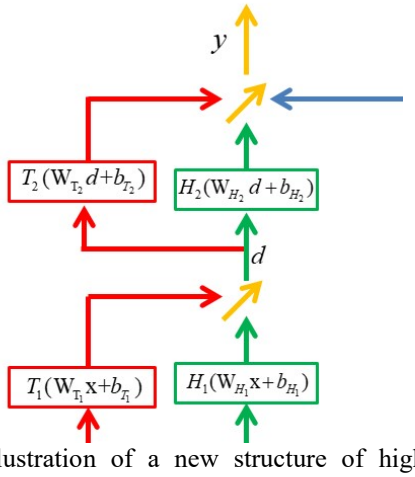

Fig. 2 Illustration of a new structure of highway networks

where $\mathrm{x}, y \in \mathrm{R}^m, d \in \mathrm{R}^n$ denote the input, output and middle input of the whole layer, respectively, $\mathrm{W}_{T_1}, \mathrm{W}_{H_1} \in \mathrm{R}^{n\times m}, b_{H_1}, b_{T_1} \in \mathrm{R}^n, \mathrm{W}_{T_2}, \mathrm{W}_{H_2} \in \mathrm{R}^{m\times n}, b_{H_2}, b_{T_2} \in \mathrm{R}^m$ denote the weights and bias in the first layer and the second layer respectively and H_1, H_2, T_1, T_2 are the activation function. Here, x and y are called pass units and d indicates growing units.

This structure provides some advantages. According to the values of $T_1(\mathrm{W}_{T_1}\mathrm{x}+b_{T_1})$, it can be determined whether or not the number of neurons in growing units is sufficient. If the values are large enough, this shows that the utility of $H_1(\mathrm{W}_{H_1}\mathrm{x}+\mathrm{b}_{H_1})$ is greater than the utility of 0. In consequence, the dimensionality of growing units, n, can be raised until the values of $T_1(\mathrm{W}_{T_1}\mathrm{x}+b_{T_1})$ are small enough or achieve the maximum number of growing units permitted. As a result, the dimensionality can change dynamically according to the utility of $H_1(\mathrm{W}_{H_1}\mathrm{x}+\mathrm{b}_{H_1})$. It is worth noting that the values of $T_1(\mathrm{W}_{T_1}\mathrm{x}+b_{T_1})$ is related to the input. Therefore, the input should be selected suitably to get representative value to check a reasonable width in each layer. However, the structure is too complex to be trained. Figure 3 shows the result of CIFAR 10 dataset.

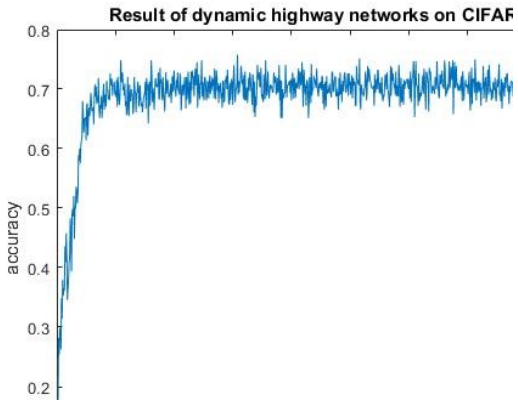

Fig. 3 Result of dynamic highway networks on CIFAR10

From Fig. 3, the performance on CIFAR10 is not high enough. As a result, a new structure of dynamic highway networks is designed.

$$d = T_1(\mathrm{b}_{T_1})\cdot H_1(\mathrm{W}_{H_1}\mathrm{x}+\mathrm{b}_{H_1}), \tag{9}$$

$$y = T_2(b_{T_2})\cdot H_2(\mathrm{W}_{H_2}\,d+b_{H_2}) + (1-T_2(b_{T_2}))\cdot \mathrm{x}, \tag{10}$$

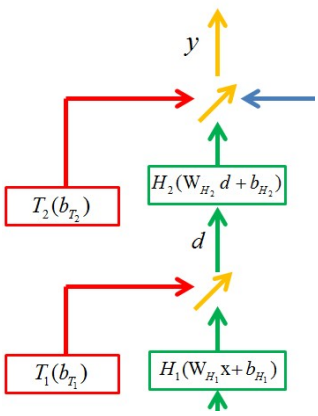

Fig. 4 Illustration of dynamic highway networks

where $\mathrm{x}, y \in \mathrm{R}^m, d \in \mathrm{R}^n$ denote the input, output and middle input of the whole layer respectively, $\mathrm{W}_{H_1} \in \mathrm{R}^{n\times m}, b_{H_1}, b_{T_1} \in \mathrm{R}^n, \mathrm{W}_{H_2} \in \mathrm{R}^{m\times n}, b_{H_2}, b_{T_1} \in \mathrm{R}^m$ denote the weights and bias in the first layer and the second layer, respectively and H_1, H_2, T_1, T_2 are the activation function. Here, x and y are called pass units and *d* indicates growing units.

This structure has some advantages. First, the outputs of T_1 and T_2 are independent of the input. The original highway networks shows that the activity of the transform $T(\mathrm{W}_T\mathrm{x}+b_T)$ varies with different inputs (Srivastava et al., 2015). It is flexible because the transform $T(\mathrm{W}_T\mathrm{x}+b_T)$ can compare the utility of input x and the output of transform $H(\mathrm{W}_H\mathrm{x}+b_H)$ with different input x. Therefore, the new structure helps to understand the utility of $H(\mathrm{W}_H\mathrm{x}+b_H)$ for all inputs. In order to verify that, ignoring x will give small change. We test the difference between transform $T(\mathrm{W}_T\mathrm{x}+b_T)$ and transform $T(b_T)$, and the results are displayed in Figure 5.

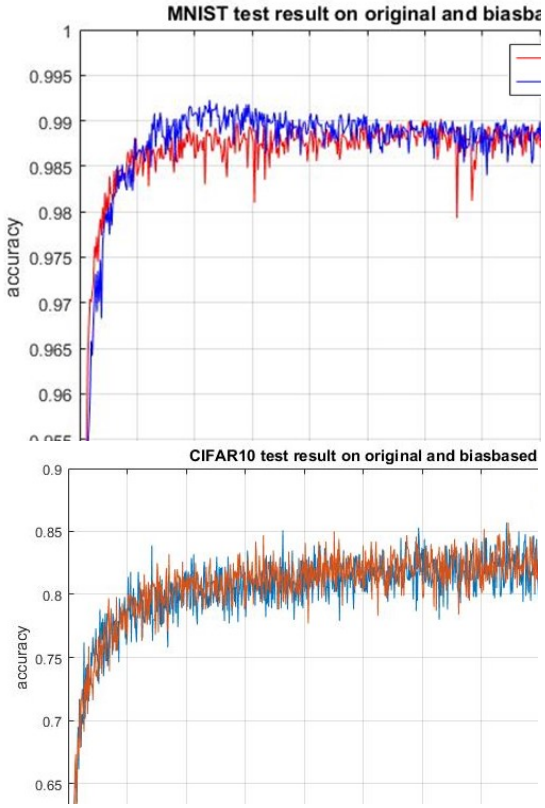

Fig. 5 Result on original and bias-based highway networks

In Fig. 5, it can be verified that ignoring the input x does not cause a huge deviation from the original highway networks.

In addition, there are two sublayers in the new structure. The fundamental concept of two sublayers is similar to that used in residual networks (Wu et al., 2017). It is said that "*if the residual function has only a single layer, the transformation function is similar to a linear layer, for which we have not observed advantages.*" Thus, there are some benefits in the two-sublayer structure. In dynamic highway networks, the function of the first layer is as shown in equation (9). In this layer, $T_1(b_{T_1})$ compares the utility of $H_1(\mathrm{W}_{H_1}\mathrm{x}+b_{H_1})$ and 0.

$T_1(b_{T_1})$ replaces the $T_1(\mathrm{W}_{T_1}\mathrm{x}+b_{T_1})$ as the determination of whether the number of neurons is sufficient. And the characteristics of $T_1(b_{T_1})$ are similar to $T_1(\mathrm{W}_{T_1}\mathrm{x}+b_{T_1})$ as mentioned previously. Consequently, the dimensionality can change dynamically according to the utility of $H_1(\mathrm{W}_{H_1}\mathrm{x}+b_{H_1})$. The added growing units are initialized by a Gaussian distribution function with zero mean and 0.01 variance. Therefore, the added weights are small enough to avoid the performance in the training step shuddering. The second layer is designed to correct the dimensionality of y, which should be the same as the dimensionality of the pass units. $T_2(b_{T_2})$ can compare the utility of $H_2(\mathrm{W}_{H_2}\mathrm{x}+b_{H_2})$ and input x such that the goodness of original highway networks can be preserved.

There are some parameters that should be determined before training. First, the number of layers should be decided, as this influences the ability to deal with the abstract. It enhances the capacity for learning complex data in deeper networks. Next, the width of the pass units should also be specified as this affects the quantity of information that can be passed, regardless of the width of the growing units. No matter what information can be carried in growing units, it still needs to be reduced to the dimensionality of the pass units. Therefore, the number of pass units plays an important role in dynamic highway networks. Finally, the dimensionality of the growing units should be initialized. Although it will change dynamically through the training step, it influences the time required for training and should be determined.

According to equation (9) and equation (10), the number of parameters is 2*m*n+2*n+2*m. In comparison, the number of parameters in the original highway networks is 2*m*n+2*n. There is thus no huge difference between the numbers of parameters in these two structures. However, this new structure provides the ability to change the number of neurons automatically.

To obtain a better result on $T_1(b_{T_1})$ and $T_2(b_{T_2})$, a regularization term is added. Figure 6 shows the regularization function used.

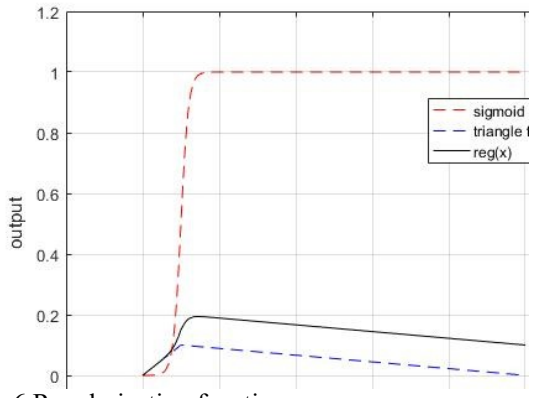

Fig. 6 Regularization function

Here the red line is a sigmoid function, the blue line is a triangle function and the black line which is the regularization function used is a linear combination of the red line and blue line and is shown in equation (11).

$$reg(\mathrm{x}) = \begin{cases} 0.1 * \dfrac{1}{1+e^{-100(\mathrm{x}-0.1)}} + x, & x \le 0.1 \\ 0.1 * \dfrac{1}{1+e^{-100(\mathrm{x}-0.1)}} + \dfrac{1-x}{9}, & x > 0.1 \end{cases}, \quad (11)$$

By using this regularization function, $T_1(b_{T_1})$ and $T_2(b_{T_2})$ tend to be close to 0 or 1 in order to reduce loss. Therefore, it can emphasize the utility of $H_1(\mathrm{W}_{H_1}\mathrm{x}+b_{H_1})$ and $H_2(\mathrm{W}_{H_2}\mathrm{x}+b_{H_2})$. Additionally, the regularization term for weights and dropout is used to avoid overfitting. The loss is calculated by cross-entropy and the regularization for weights here is calculated by the L2-norm. Equation (12) illustrates the entire loss function.

$$\begin{aligned} Loss = &-\sum_{x}\left[y(\mathrm{x})\log(\hat{y}(\mathrm{x})) + (1-y(\mathrm{x}))\log(1-\hat{y}(\mathrm{x}))\right] \\ &+\lambda_T\left[\sum_{l}\sum reg(T_{l,1}(\mathrm{b}_{T_1})) + \sum reg(T_{l,2}(\mathrm{b}_{T_2}))\right] + \lambda_W \|W\|_2, \end{aligned} \quad (12)$$

where $y(\mathrm{x})$ and $\hat{y}(\mathrm{x})$ are the ground truth and prediction of data x, the parameters λ_T and λ_W control the importance of the regularization term; the first regularization term has been described above and it regularizes $T_1(b_{T_1})$ and $T_2(b_{T_2})$, the second regularization term minimizes all the weights used in the whole network. Adam (Kingma and Ba, 2015) is applied to optimize the weights in the structure.

Convolution neural networks based dynamic highway networks

In order to increase the performance of dynamic highway networks on image testing, it is helpful to use convolutional neural networks (CNN). The primary difference between convolutional neural networks and neural networks is shared weights. Because the number of neural networks needed to deal with numerous inputs is too vast to train, convolutional neural networks define shared weights, called filters, to reduce the parameters. The filters convolute the inputs and get outputs that are called feature maps. Figure 7 illustrates the shared weights, where lines in the same color denote the same weights.

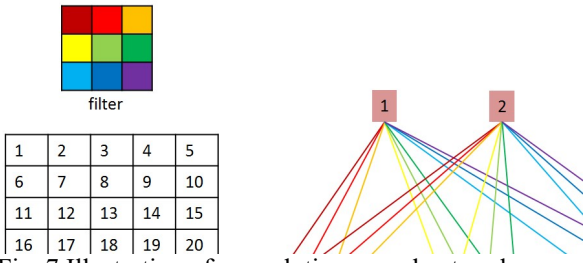

Fig. 7 Illustration of convolution neural networks

Therefore, CNN-based dynamic highway networks are designed. W_{H_1} and W_{H_2} in equation (9) and equation (10) are replaced by two sets of filters. The gates $T_1(b_{T_1})$ and $T_2(b_{T_2})$ are two sets which contain gates. Each gate is a single value and controls a corresponding feature map. The number of feature maps is the same implication of the number of neurons in original dynamic highway networks. Thus, they can also be divided into growing feature maps and pass feature maps, and give the same meaning of growing units and passing units in original dynamic highway networks. The mechanism for changing the dimensionality of growing feature maps dynamically depends on the value of gates $T_1(b_{T_1})$ as well and Figure 8 illustrate the structure.

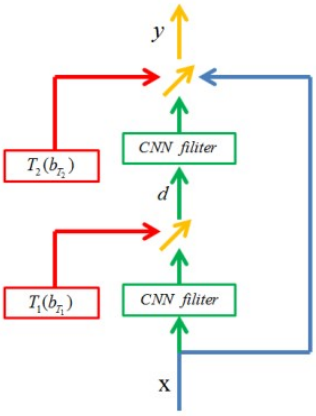

Fig. 8 Illustration of CNN-based dynamic highway networks

EXPERIMENTS FOR CNN-BASED DYNAMIC HIGHWAY NETWORKS

To make sure that the dynamic CNN-based highway network is effective, it is tested on the MNIST dataset and the result is shown in Figure 9.

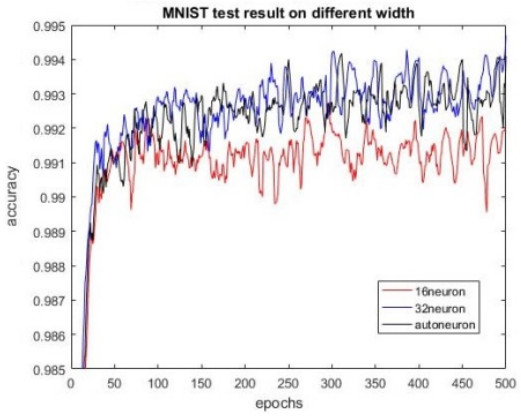

Fig. 9 Result on different widths on MNIST

Here there are three networks in 10 layers and different widths. The red and blue lines are the structure in Fig. 4, but the dimensionality of the growing units is kept at 16 and 32 respectively and cannot be changed. The black line is the structure in Fig. 4 and the dimensionality of the growing units can be changed dynamically, and it is called "autoneuron" here. Fig. 9 shows an obvious difference between the accuracy of the 16-neuron and 32-neuron networks. The accuracy of autoneuron networks can be similar to 32 neurons but achieved with fewer neurons. Figure 10, Figure 11 and Figure 12 display the value of $T_1(b_{T_1})$.

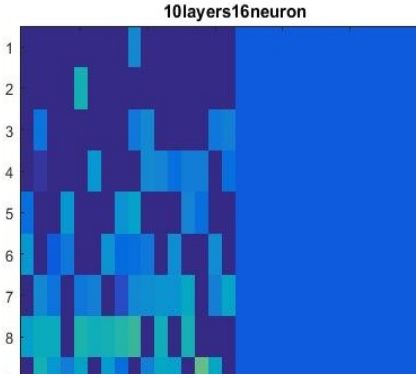

Fig. 10 $T_1(b_{T_1})$ in 16-neuron networks

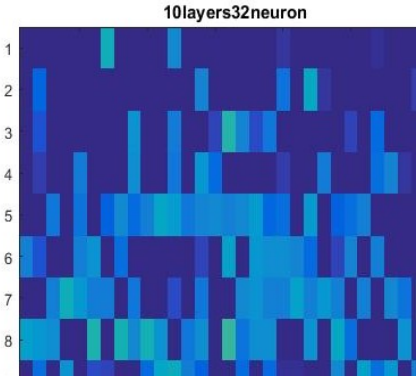

Fig. 11 $T_1(b_{T_1})$ in 32-neuron networks

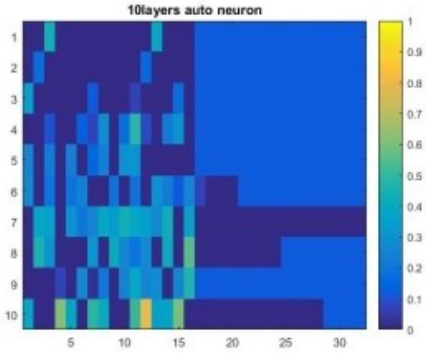

Fig. 12 $T_1(b_{T_1})$ in autoneuron networks

Table 1. Test error for different methods on MNIST

Method	Number of the neuron for each layer	Number of parameters	Test Error
Highway Network (Srivastava et al., 2015)	32	151,000	0.0045
Dynamic Highway Network	32	92,000	0.007
	64	184,000	0.006
	autoneuron	116,000	0.006

Fig. 10, Fig. 11 and Fig. 12 display the values of $T_1(b_{T_1})$ in a row in the same layer. The unused neurons are filled with the same value. Based on Fig. 10 and Fig. 12, the networks can achieve similar accuracy to that of 32-neuron networks if we just append some neurons on 16-neuron networks in specific layers. Therefore, the fact that autoneuron networks can change the dimensionality dynamically is validated.

Table 1 shown the MNIST classification results. In this examination, the number of parameters in 32-neuron networks is about 184,000 and the number of parameters in dynamic networks is about 116,000. The proposed structure reduces the parameters by 36%. The dynamic network reduces about 23% parameters than the highway network and the error rates are very close to the highway network. As a result, adjusting the dimensionality manually is not necessary to enhance the performance with the mechanism proposed. $T_1(b_{T_1})$ can help us to tune the best dimensionality with fewer neurons.

Furthermore, the proposed structure is further tested on CIFAR10 dataset to make sure the performance of CNN-based dynamic highway network. The result is shown in Figure 13 and Table 2.

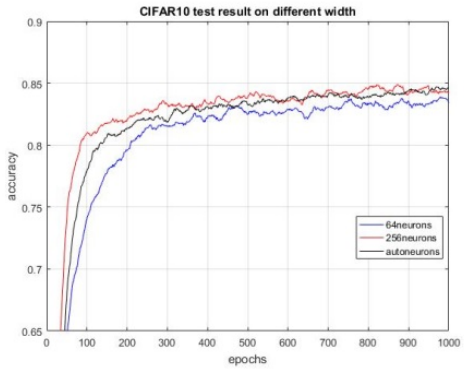

Fig. 13 Result on different widths on CIFAR10

In Figure 15, there are three networks in 20 layers and different widths and the result is processed. The red and blue lines are the structure whose dimensionality of the growing units is kept at 256 and 64, respectively and cannot be changed. The black line is autoneuron structure. Fig. 15 shows an obvious difference between the accuracy of the 64-neuron and 256-neuron networks. The accuracy of autoneuron networks is similar to that of 256 neurons but with fewer neurons. Table 2 shows the detail about the number of parameters and the error rates used for different width. In Table 2, it shows that compared to 256 neuron numbers. Autoneuron can reduce the number of neurons about 45% and does not reduce the test error. This also means that there are redundant neurons in the model.

Table 2. Test error for different methods on cifar10

Method	Number of the neuron for each layer	Number of parameters	Test Error
Highway Network (Srivastava et al., 2015)	64	2,300,000	0.0754
Dynamic Highway Network	64	368,000	0.1612
	256	1,472,000	0.1519
	autoneuron	864,800	0.1518

The details are shown in Figure 14, Figure 15 and Figure 16. The values of $T_1(b_{T_1})$ are shown there.

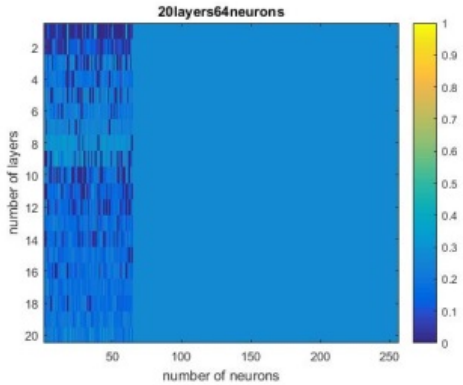

Fig. 14 $T_1(b_{T_1})$ in 64-neuron networks

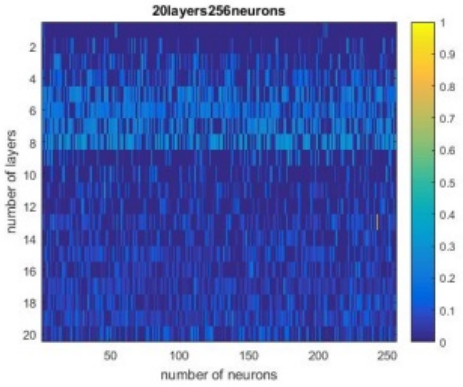

Fig. 15 $T_1(b_{T_1})$ in 256-neuron networks

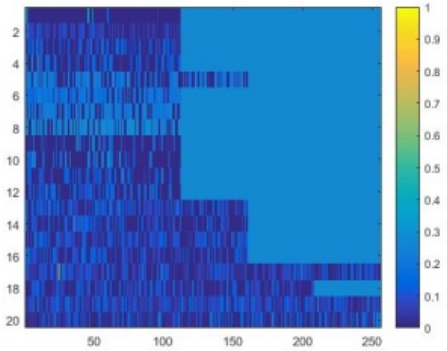

Fig. 16 $T_1(b_{T_1})$ in autoneuron networks

Fig. 14, Fig. 15 and Fig. 16 display the values of $T_1(b_{T_1})$ in a row in the same layer. The unused neurons are filled with the same value. Based on Fig. 16 and Fig. 18, the networks can achieve similar accuracy to that of 256-neuron networks if we just append some neurons on 64-neuron networks in each layer. Therefore, the fact that autoneuron networks can change the dimensionality dynamically is further validated. By automatically adjusting method, we can avoid the step of selecting the number of neurons, and save a lot of trial and error time. At the same time, our method can train the model structure that is most suitable for the application.

CONCLUSION

In this paper, the CNN-based dynamic highway network is proposed. Through the characteristics of highway networks, the utility of neurons in each layer can be checked. By appending neurons to the proper layer, it can raise the performance as high as complex networks. Besides, a regularization function is designed to tune the suitable value of gates so that the utility of neurons can be determined more precisely. Two picture datasets, MNIST and CIFAR10, are used to justify the performance of dynamic highway networks. From the result, the proposed structure can achieve high performance in less parameter.

REFERENCES

Bengio, Y., Simard, P., and Frasconi, P., "Learning long-term dependencies with gradient descent is difficult," *IEEE Transactions on Neural Netwroks, 5*(2), 157-166 (1994).

Guo, Y., Liu, Y., Oerlemans, A., Lao, S., Wu, S., and Lew, M. S., "Deep learning for visual understanding: a review," *Neurocomputing, 187*, 27-48 (2016).

He, K., Zhang, X., Ren, S., and Sun, J., "Deep residual learning for image recognition," *Proceeding of the IEEE conference on computer vision and pattern recognition*, 770-778 (2016).

Kingma, D. P., and Ba, J., "Adam: a method for stochastic optimization," *Proceeding of International Conference on Learning Representations*, San Diego, CA, 1-15 (2015).

Krizhevsky, A., Sutskever, I., and Hinton, G. E., "ImageNet classification with deep convolutional neural networks," *Proceeding of 25th International Conference on Neural Information Processing System*, 1079-1105 (2012).

Li, K., Bare, B., Yan, B., Feng, B., and Yao, C., "HNSR: highway networks based deep convolutional neural networks model for single image super-resolution," *Proceeding of IEEE International Conference on Acoustics, Speech and Signal Processing*, Calgary, AB, Canada, 1478-1482 (2018).

Lin, Z., Ji, K., Kang, M., Leng, X., and Zou, H., "Deep convolutional highway unit network for SAR target classification with limited labeled training data," *IEEE Geoscience and Remote Sensing Letters, 14*(7), 1091-1095 (2017).

Lu, L., and Renals, S., "Small-footprint highway deep neural networks for speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing, 25*(7), 1502-1511 (2017).

Oyedotun, O. K., Shabayek, A. E. R., Aouada, D., and Ottersten, B. o., "Highway network block with gates constraints for training very deep networks," *Proceeding of IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, Utah, U.S. (2018).

Srivastava, R. K., Greff, K., and Schmidhuber, J., "Training very deep networks," *Proceeding of 28th International Conference on Neural Information Processing Systems*, Montreal, Canada, 2377-2385 (2015).

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A., "Going deeper with convolutions," *Proceeding of IEEE Conference on Computer Vision and Pattern Recognition* Boston, MA, USA, 1-9 (2015).

Wang, X., Yu, F., Dou, Z.-Y., and Gonzalez, J. E., "Skipnet: Learning dynamic routing in convolutional networks," *arXiv preprint arXiv:1711.09485* (2017).

Wu, S., Zhong, S., and Liu, Y., "Deep residual learning for image steganalysis," *Multimedia Tools and Application, 77*(9), 10437-10453 (2017).

Zagoruyko, S., and Komodakis, N., "Wide residual networks," *arXiv preprint arXiv:1605.07146* (2016).

NOMENCLATURE

$\mathbf{x}$ the input vector of the layer

y the output vector of the layer

$y(x)$ the ground truth label of data x

$\hat{y}(x)$ the prediction label of data x

W_H the weight of the highway network

b_H the bias of the highway network

λ_W the penalty weight of the highway network term

λ_T the penalty weight for the gate term

H the sigmoid activation function for the weight

T the sigmoid activation function for the gate

L the loss function

W_{H_1} the weight of highway network in the first layer

W_{H_2} the weight of highway network in the second layer

W_{T_1} the weight of gate in the first layer

W_{T_2} the weight of gate in the second layer

b_{H_1} the bias of highway network in the first layer

b_{H_2} the bias of highway network in the second layer

b_{T_1} the bias of the gate in the first layer

b_{T_2} the weight of the gate in the second layer

H_1 the sigmoid activation function for the highway network in the first layer

H_2 the sigmoid activation function for the highway network in the second layer

T_1 the sigmoid activation function for the gate in the first layer

T_2 the sigmoid activation function for the gate in the second layer

$reg(x)$ the regularization function

基於動態高速公路類神經網路之圖像識別研究

王隨賢　林威志　黃漢邦
國立台灣大學機械工程學系

摘 要

基於深度學習的發展，許多成功的機器學習模型都相對的巨大。並且需要使用多次的實驗來測試出比較適合的類神經網路結構，此步驟不僅僅耗時也可能找出具有冗餘神經元的架構。為此本文旨在發展一套新式類神經架構，稱為動態高速公路類神經網路。其架構具備高速公路的特性能夠在訓練過程中藉著調整閥門數值，使部分輸入資料傳遞至輸出部分。同時此架構也具備動態調整神經元個數的功能，從較小的類神經網路架構在各層逐漸加入神經元，使得整體類神經模型能夠達到最佳性能。並以 MNIST 與 CIFAR10 兩種數據驗證此架構之性能。