DCH-Net: Densely Connected Highway Convolution Neural Network for Environmental Sound Classification

Xiaohu Zhang，Yuexian Zou*
ADSPLAB, Peking University Shenzhen Graduate School, Shenzhen, China
*corresponding author {zouyx@pkusz.edu.cn}

***Abstract*—Environmental Sound Classification (ESC) plays a vital role in the field of machine auditory scene. Recently, the Highway Network CNN model has achieved the state-of-art results via solving the vanishing-gradient problem of much deeper CNN. However, carefully analyzing the Highway Network model shows that the Highway Network model lacks ability to maximize information flow between layers, which is essentially benefits the discriminative representation of acoustic events. Besides, the Highway Network model size is larger than 20MB for ESC task, which is still large for mobile applications. Regarding to these two issues, in this study, we propose a novel Densely Connected Highway Convolutional Network (DCH-Net) model for ESC task. Specifically, a densely highway module is developed which is able to ensure the maximum information flow between layers by connecting all layers directly with each other. Besides, to reduce the model size, a global average pooling layer is designed which replaces the traditional fully connection layers and the parameters of the model is greatly reduced. Experimental results show that our DCH-Net ESC model achieves accuracy of 69% and 90% on ESC50 and ESC10 dataset respectively, which is 2% and 10% higher than that of Highway Network based Highway networks ESC model. Meanwhile our model size is only 2MB.**

Keywords—Convolution Neural Network; Environmental Sounds Classification; Densely Connected Highway module; Global Average Pooling

I. INTRODUCTION

Convolutional neural networks (CNNs) have achieved many state-of-art results in environmental sound classification (ESC) tasks [1, 2, 3]. In the last few years, there are plenty of improvements on CNN-based models. Specifically, very deep CNNs like VGG have shown impressive results [4] since hierarchical convolution max-pooling layers have the ability to extract very abstract high-level features.

As CNNs become increasingly deep, a new research problem emerges: its gradient is easy to be vanished during the training process. To solve the gradient vanishing problem, some variants of deep neural networks have been proposed. For example, Highway Networks [5,6] have been proposed to alleviate the vanishing gradient problem through bypassing signal from earlier layer to the later via identity connections. This identity connections allows better information and gradient flow. FractalNets [17] randomly drop layers during training to reduce the depth of network. For ESC task, Boddapati [17] proposed a 22 layer GoogleNet, a specifically designed highway network. He has evaluated the model on ESC-50 and ESC-10 datasets, respectively. Experimental results show that GoogleNet model performs much better than traditional VGG model. It is noted that these approaches share a common characteristic: short paths are created for feature information and gradient flow.

Although Highway Network proposed above is effective to alleviate the vanishing gradient problem, but it cannot ensure the maximum information flow between layers. Therefore, for the purpose of doing that, we decide to seek for new methods. In image classification task, Huang [7] proposed a densely connected highway neural networks to maximize information flow between layers. In his design, instead of only connecting two layers of a network, he selects to connect all layers directly with each other. Thus, this architecture has largely increased the flow of information and gradients throughout the network. Meanwhile, a possible effect of this dense connectivity pattern is that it would not relearn redundant feature maps which leads to much less redundant parameters.

Carefully examining the sound event classification task, we find that sound of a specific acoustic event may be produced by a wide variety of sources, which is much more transformative than human speech, it is expected that the increasing of flowing of feature information is essentially benefit the discriminative representation of acoustic events. Based on this assumption, in this paper, we design to use the densely connected highway neural networks for the sound event classification task. Considering that a relatively small model is needed in mobile devices. Traditionally, environmental sounds have weak absolute locality in the time-frequency spectrogram [8]. Hence, it may be inferred that the spatial information extracted in high-level feature maps does not contribute much to the final classification accuracy. Therefore, we try to reduce the model size through reducing the dimension of these spectrograms. A global average pooling layer has been proposed in NIN model [9] to substitute all traditional fully connected layers. The main principle behind the idea is that global average pooling layer only remains the statistical feature in feature maps. Thus, by applying global average pooling to substitute all fully connected layers in convolution networks, only the global features of sound event are retained.

II. The Proposed Method

In this section, we present the scheme of the proposed Densely Connected Highway Convolution Neural Network (DCH-Net), whose architecture is shown in Figure. 1.

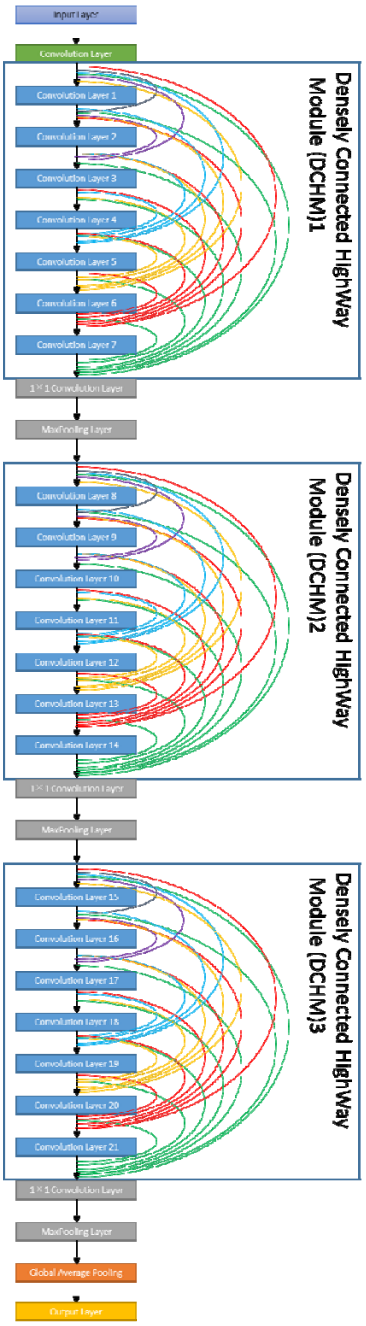

Figure 1 Architecture of the proposed Densely Connected Highway Convolution Neural Network (DCH-Net)

As shown in Figure 1, following our design in [7], DCH-Net adopts two channel input structure which takes the log-mel acoustic feature and delta acoustic feature as input, respectively. The log mel feature represents the static characteristic of sound event while the delta feature represents the dynamic characteristic of sound event. For the Convolution Layer (green block), 16 filters with the size of 3×3 are used. It is clear that DCH-Net mainly consists of three Densely Connected Highway Modules (DCHMs). For each DCHM, there are 7 convolution layers (marked as convolution layer 1 to 7). In our design, the filter size is of (3×3) for all filters for its simplicity. But the number of filters used in each convolution layer is different which is determined by the formula 4+N×12. It is noted here N is the index number of the convolution layer. As shown in Figure 1, taking the convolution layer 3 as one example, N is set to 3. In this case, the number of filters used in the convolution layer 3 is 40. The details of designing DCHM is given in Sect. 2A. Besides, from Figure 1, we can see that, after each DCHM, a convolution layer with 1×1 filter size and the number of convolution filters is set as the same as that of the previous convolution layer and a max-pooling layer with 2×2 pooling size have been designed. The 1×1 convolution filter is used to increase the ability of non-linear fitting of the network and the max-pooling layer is used to extract more abstract high-level features. At last, aiming at reducing parameters, a global average pooling is used to substitute the traditional fully connected layers and a softmax classifier is adopted to produce the classification results.

It is noted that, in Highway networks [7], ReLU activation function usually has been used. However, it is well known that ReLU maps the negative input to zero which would cause the loss of negative feature information of sound events. In our DCH-Net, Swish activation function [19] is used to substitute ReLU activation function since Swish activation function uses a non-linear function to compress all negative input which partially remains the negative features.

A. *The Densely Connected Highway Module*

Traditionally, in highway module, the output of a proceeding layer is connected to the output of a previous layer by using bypassing path. This bypassing path not only alleviates the vanishing-gradient problem, but also enables the feature information flow between layers. Making use of this structure, aiming at maximizing the feature information flow, Densely Connected Highway Module (DCHM) is designed which is essentially an extension of the traditional highway module. To better understand the structures of the Highway module and DCHM, the comparison of them is presented in Figure 2. To make things clear, for the convolution layer l, its input feature and output feature are denoted as x_{l-1} and y_l , respectively.

As shown in Figure 2(a), there are only three bypassing paths in highway module. Every bypassing is used to add the input feature of each convolution layer to its output feature. The sum result of each layer can be calculated as follows:

$$x_{l-2}=y_{l-2}+x_{l-3},\ x_{l-1}=y_{l-1}+x_{l-2},$$

$$x_l=y_l+x_{l-1}=y_l+y_{l-1}+y_{l-2}+x_{l-3} \quad (1)$$

From Figure 2(b), we can see that DCHM uses the bypassing paths to transmit the input feature of one convolution layer to all convolution layers followed it. As a result, for the convolution layer l, the resulting input feature of next convolution layer is calculated as follows:

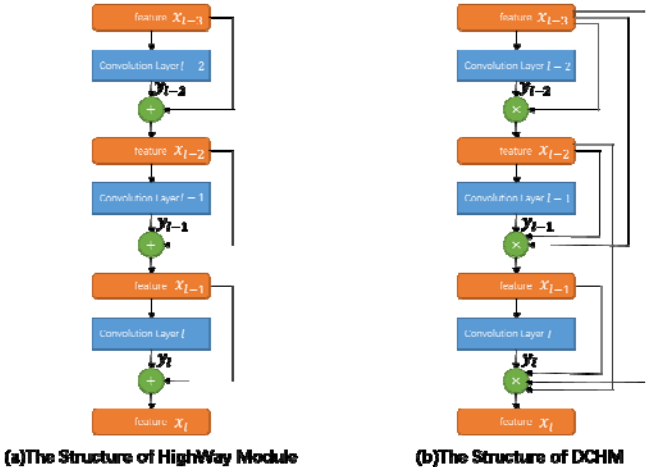

Figure 2 Structure comparison of Highway Module and DCHM

$$x_{l-2}=[y_{l-2}, x_{l-3}],\ x_{l-1}=[y_{l-1}, x_{l-2}, x_{l-3}],$$

$$x_l=[y_l, x_{l-1}, x_{l-2}, x_{l-3}] \qquad (2)$$

where the symbol [.] represents the concatenation operation. Compared (2) with (1), we can see that DCHM concatenates features while Highway module adds features, It is also clear that the dimension of feature maps generated by DCHM is higher than that of the Highway module. In addition, since the DCHM concatenates the input feature of all previous layers together, feature maps generated by DCHM would contain much more feature information than that of the Highway module. Figure 3 gives the comparison of high level features extracted by highway module and DCHM. The dots are obtained by T-SNE projection of the feature vectors obtained at the output of the last hidden layer by the highway module and DCHM, respectively. For illustration purpose, Figure 3 plots the feature vectors generated by the DCHM and the highway module, respectively. It is easily seen that features extracted by the DCHM is much more discriminative than that of the Highway module. These results indirectly demonstrate the high-level feature extraction capability using the concatenated features from the bypassing paths.

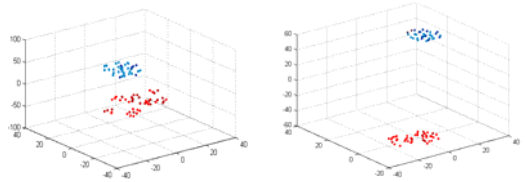

Figure 3 the illustration of high-level features of two classes extracted by highway module and DCHM

B. The Global Average Pooling

As shown in Figure 1, in our design, the global averaging pooling is adopted. The main purpose is to further reduce the model size since the model size of highway network is still large, which is not suitable when the computation ability is limited. Carefully analyzing shows that the large model size mainly is due to its fully connected layers in the Highway networks. It is a common practice to use the global average pooling layer to replace the fully connection layers. In this section, we describe the global average pooling through a comparison of using fully connected layer in DCH-Net or global average pooling in DCH-Net. Figure 4 (a) shows the structure of a fully connected layers used in the DCH-Net. For explicit explanation, following parameters are defined: N_a is the number of feature map generated by the highway module. For each feature map, the length is defined as T and the width is defined as R. Additionally, for the fully connected layer FCL1 and FCL2 in DCH-Net, the number of neurons is defined as N_{fc}, and the number of neurons in the output layer is defined as N_o.

From Figure 4(a), we could see that the N_a feature maps with $T\times R$ size generated by the DCHM is transformed into a 1-dimensional vector. Then this vector is input into FCL1 directly. In addition, the second fully connected layer FCL2 is followed by the first fully connected layer FCL1. Finally, an output layer is followed by the two fully connected layers. As a result, the parameter size A_1 between the HM and output layer in Figure 4(a) can be calculated as

$$A_1 = (1\times T\times R\times N_a\times N_{fc}+N_{fc})+(N_{fc}\times N_{fc}+N_{fc})+(N_{fc}\times N_o+N_{fc}) \quad (3)$$

In order to reduce parameter size A_1, we propose a global average pooling layer to substitute all fully connected layers, which is demonstrated in Figure. 4(b).

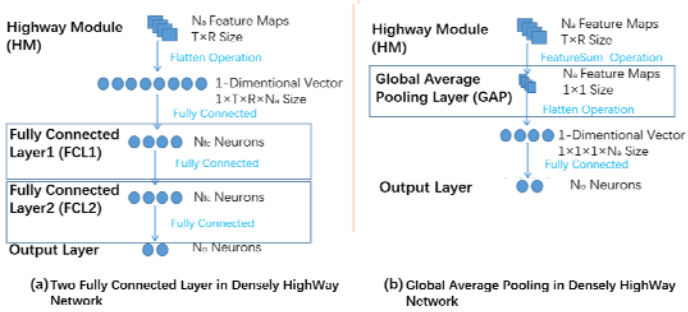

Figure 4 (a) two fully connected layers used in the DCH-Net *model; (b) a GAP layer used in our proposed DCH-Net model.*

As shown in Figure. 4 (b), a global average pooling layer (GAP) is followed by the Highway Module (HM) and the output layer is directly followed after the global average pooling layer (GAP). Specifically, the GAP makes an average pooling operation for each feature map generated from HM, which described in detail in Figure. 5.

In Figure. 5, the output of GAP is denoted as b_i which is the average value of each feature a_{ix} in the i-th feature map. This design is triggered by the experimental findings where, for environmental sounds, there is weak absolute locality in the time-frequency spectrogram. Therefore, using more spatial high-level feature map would have less benefit on improving the final classification accuracy. Therefore, based on this observation, the global average pooling calculates the statistic values of features and thus renders more abstract spatial feature maps. As shown in Figure. 4 (b), the parameter size A_2

between the HM and output layer in Figure 4(b) is calculated by eqn. (4):

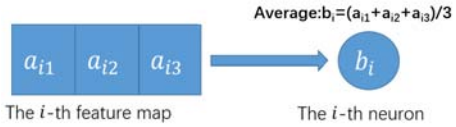

Figure 5. Illustration of the Global Average Pooling (GAP) Operation (a_{ij} represents the j-th value in the i-th feature map, b_i represents the value of the i-th neuron in the GAP Layer).

$$A_2 = (1 \times 1 \times N_a \times N_o) + N_a \quad (4)$$

Comparing (3) with (4), we could easily find that A_2 is much smaller than A_1.

To validate the effectiveness of using GAP instead of fully connected layers, we intuitively compare the feature distribution of one ESC task by the T-SNE visualization tool. The results are presented in Figure. 6. It is clear to see that the feature distributions with GAP or without GAP are changed slightly, which may give a similar discriminability for environmental sound classification task.

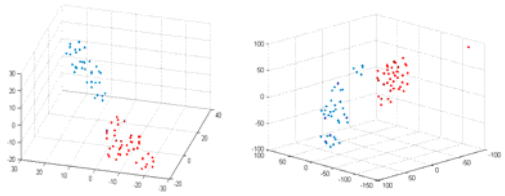

Figure 6. Illustration of features maps output by the output layer with and without the global average pooling operation.

III. Experiments and Results

To our knowledge, this is a first try for developing DCH-Net-ESC system for ESC task. Therefore, conducting performance evaluation is one of our main tasks in this work. In the following, the details of datasets used in experiments are introduced firstly. Then, the experimental settings are given in subsection B. Finally, we evaluate the performance of our DCH-Net-ESC system.

A. Datasets

We use two datasets (ESC50 and ESC10) to evaluate our proposed DCH-Net. Detail information about these two datasets are described in Table I.

TABLE I. Basic Information of Datasets

Datasets	Classes	Train/Test	Duration
ESC50	50	80%/20%	2.8 hours
ESC10	10	80%/20%	0.6 hours

As shown in Table I, The ESC50 dataset consists 50 classes of sound event. All sounds in ESC50 is about environmental sound event， e.g., birds singing, dog barking, raining and so on. The ESC10 dataset contains all about human behavior sounds. It is noticeable that both the ESC50 and ESC10 are recorded in real environment. For both datasets, we choose 80% of audios for training and 20% of audios for testing.

B. Performance Comparison

For the purpose of evaluating the performance of our system, we use several basic ESC systems as our baseline, which are list as follows:

1) KNN-ESC: MFCC feature with KNN classifier [15]
2) SVM-ESC: MFCC feature with SVM classifier [15]
3) RamdomForest-ESC: MFCC feature with RamdomForest classifier [15]
4) multi-kernelSVM-ESC: MFCC feature with different kernel SVM classifiers fused together [16]
5) MFCC-AlexNet-ESC: MFCC feature with AlexNet classifier [17]
6) MFCC-GoogleNet-ESC: MFCC feature with GoogleNet classifier [17]
7) Spectrograms-CRNN-ESC: STFT Spectrograms feature with GoogleNet classifier [17]
8) PiczakCNN-ESC: log mel-spectrum and delta-spectrum feature with 7-layer CNN classifier [10]
9) auDeepRNN-ESC: MFCC feature with Deep RNN classifier [18]
10) Mel-Spectrogram-CLNN-ESC: log Mel-Spectrogram feature with CLNN classifier [14]
11) Mel-Spectrogram-MCLNN-ESC: log Mel-Spectrogram feature with MCLNN classifier [14]
12) Spectrograms-AlexNet-ESC: STFT Spectrograms feature with AlexNet classifier [17]
13) Spectrograms-GoogleNet-ESC: STFT Spectrograms feature with GoogleNet classifier [17]
14) Our Proposed DCH-Net-ESC: log Mel-Spectrogram feature with DCH-Net classifier

For our DCH-Net, SGD algorithm is used for model training and the cross-entropy is used as the loss function. The experimental results are shown in Table Ⅱ.

TABLE Ⅱ. Performance Comparison of Different ESC Systems

ESC system	ESC10	ESC50	Model Size
KNN-ESC [15]	66.7%	32.2%	-
SVM-ESC [15]	67.5%	39.6%	-
RamdomForest-ESC [15]	72.7%	44.3%	-
multikernelSVM-ESC [16]	-	62.2%	-
MFCC-AlexNet-ESC [17]	73.0%	44.9%	60M
MFCC-GoogleNet-ESC [17]	75.9%	49.1%	20M
Spectrograms-CRNN-ESC [17]	-	60.3%	
PiczakCNN-ESC [10]	80.3%	64.5%	105M
auDeepRNN-ESC [18]	82.7%	64.3%	-
Mel-Spectrogram-CLNN-ESC [14]	77.5%	-	-
Mel-Spectrogram-MCLNN-ESC [14]	85.5%	-	-
Spectrograms-AlexNet-ESC [17]	78.4%	63.2%	60M
Spectrograms- GoogleNet-ESC [17]	78.7%	67.8%	20M
Our Proposed DCH-Net-ESC	**90%**	**69.0%**	**2M**

From Table Ⅱ, we can see that, compared with KNN, SVM, RamdomForest and multikernelSVM, the CNN-based ESC method achieve much better accuracy. The reason is that features extracted by the deep learning-based model is more discriminative than handcrafted features. Compared with AlexNet, CRNN, PiczarkCNN, CLNN and MCLNN, GoogleNet achieves much better accuracy on ESC50, which

demonstrates that after alleviating the vanishing-gradient problem, Highway networks has better ability for high level feature extraction than that of the shallow CNNs. Compared with GoogleNet-ESC, our proposed DCH-Net gets the state-of-art results on both ESC10 and ESC50, which demonstrates that our designed DCMH has even better capability of extracting discriminative high-level features by maximizing information flowing using concatenation bypass information. Besides, with the GPA, it is encouraging to see that our DCH-Net has only about 2M model size which is much smaller than that of other models.

CONCLUSIONS

In this paper, we proposed a Densely Connected Convolution Neural Network (DCH-Net) for environmental sound classification task. In DCH-Net, a densely connected highway module is designed to improve network's ability of high level feature extraction. On the other hand, the global average pooling layer is used to substitute all fully connected layer for reducing network's parameters. Experiments show that our proposed DCH-Net ESC system achieves the state-of-art classification accuracy and much smaller model size on ESC10 and ESC50 compared to other state-of-the-art ESC methods. Obviously, the model size of Highway networks is about 20M, which is 10 times larger than our model meanwhile its classification accuracy is 2%-10% worse than our proposed model. Our future research would focus on improving classification accuracy of the lightweight convolutional neural network-based ESC system.

Acknowledge

This paper was partially supported by Shenzhen Science & Technology Fundamental Research Programs (No: JCYJ20170817160058246, JCYJ20170306165153653) & Shenzhen Key Laboratory for Intelligent Multimedia and Virtual Reality (ZDSYS201703031405467)

References

[1] Zhang, Haomin, Ian McLoughlin, and Yan Song. "Robust sound event recognition using convolutional neural networks." Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on. IEEE, 2015.

[2] Parascandolo, Giambattista, et al. "Convolutional recurrent neural networks for polyphonic sound event detection." IEEE/ACM Transactions on Audio, Speech, and Language Processing 25.6 (2017): 1291-1303.

[3] Hershey, Shawn, et al. "CNN architectures for large-scale audio classification." Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on. IEEE, 2017.

[4] Li, Juncheng, et al. "A comparison of deep learning methods for environmental sound detection." Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on. IEEE, 2017.

[5] Huang, Gao, et al. "Deep networks with stochastic depth." European Conference on Computer Vision. Springer, Cham, 2016.

[6] He, Kaiming, et al. "Deep residual learning for image recognition." Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.

[7] Huang, Gao, et al. "Densely Connected Convolutional Networks." arXiv (2016).

[8] Zhang, Haomin, I. Mcloughlin, and Y. Song. "Robust sound event recognition using convolutional neural networks." IEEE International Conference on Acoustics, Speech and Signal Processing IEEE, 2015:559-563

[9] Lin, Min, Qiang Chen, and Shuicheng Yan. "Network in network." arXiv preprint arXiv:1312.4400 (2013).

[10] Piczak, Karol J. "Environmental sound classification with convolutional neural networks." Machine Learning for Signal Processing (MLSP), 2015 IEEE 25th International Workshop on. IEEE, 2015.

[11] Deng, Lih Yuan. "The Cross-Entropy Method: A Unified Approach to Combinatorial Optimization, Monte-Carlo Simulation and Machine Learning." Technometrics 48.1(2006):147-148.

[12] Lin, Min, Q. Chen, and S. Yan. "Network In Network." Computer Science (2013).

[13] Kwan, H. K. "Simple sigmoid-like activation function suitable for digital hardware implementation." Electronics Letters 28.15(1992):1379-1380.

[14] Medhat, Fady, D. Chesmore, and J. Robinson. Environmental Sound Recognition Using Masked Conditional Neural Networks. Advanced Data Mining and Applications. 2017.

[15] Piczak, Karol J. "ESC: Dataset for Environmental Sound Classification." ACM International Conference on Multimedia ACM, 2015:1015-1018.

[16] Kumar, Anurag, and Bhiksha Raj. "Features and kernels for audio event recognition." arXiv preprint arXiv:1607.05765 (2016).

[17] Boddapati, Venkatesh, et al. "Classifying environmental sounds using image recognition networks." Procedia Computer Science 112(2017):2048-2056.

[18] Michael Freitag, Shahin Amiriparian, Sergey Pugachevskiy, Nicholas Cummins, Björn Schuller "auDeep: Unsupervised Learning of Representations from Audio with Deep Recurrent Neural Networks" arXiv:1712.04382 (2017)

[19] Ramachandran, Prajit, Barret Zoph, and Quoc V. Le. "Swish: a Self-Gated Activation Function." arXiv preprint arXiv:1710.05941 (2017).

[20] Long, Jonathan, Evan Shelhamer, and Trevor Darrell. "Fully convolutional networks for semantic segmentation." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015.

[21] Courbariaux, Matthieu, and Y. Bengio. "BinaryNet: Training Deep Neural Networks with Weights and Activations Constrained to +1 or -1." (2016).

[22] He, Tianxing, et al. "Reshaping deep neural network for fast decoding by node-pruning." Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on. IEEE, 2014.