



HAL
open science

Convolutional modulation theory: A bridge between convolutional neural networks and signal modulation theory

Fuzhi Wu, Jiasong Wu, Youyong Kong, Chunfeng Yang, Guanyu Yang,
Huazhong Shu, Guy Carrault, Lotfi Senhadji

► To cite this version:

Fuzhi Wu, Jiasong Wu, Youyong Kong, Chunfeng Yang, Guanyu Yang, et al.. Convolutional modulation theory: A bridge between convolutional neural networks and signal modulation theory. *Neurocomputing*, 2022, 514, pp.195-215. 10.1016/j.neucom.2022.09.088 . hal-03882414

HAL Id: hal-03882414

<https://hal.science/hal-03882414>

Submitted on 21 Feb 2023

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution - NonCommercial 4.0 International License

Journal Pre-proofs

Convolutional Modulation Theory: A bridge between Convolutional Neural Networks and Signal Modulation Theory

Fuzhi Wu, Jiasong Wu, Youyong Kong, Chunfeng Yang, Guanyu Yang, Huazhong Shu, Guy Carrault, Lotfi Senhadji

PII: S0925-2312(22)01180-8
DOI: <https://doi.org/10.1016/j.neucom.2022.09.088>
Reference: NEUCOM 25679

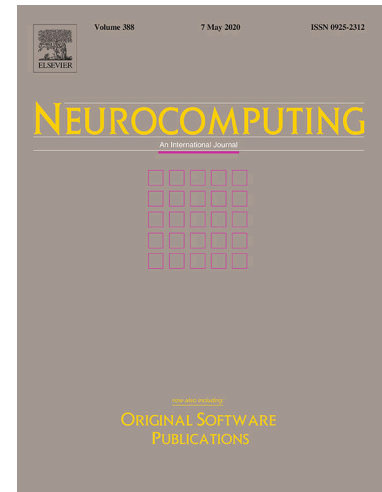
To appear in: *Neurocomputing*

Received Date: 9 August 2021
Revised Date: 5 September 2022
Accepted Date: 15 September 2022

Please cite this article as: F. Wu, J. Wu, Y. Kong, C. Yang, G. Yang, H. Shu, G. Carrault, L. Senhadji, Convolutional Modulation Theory: A bridge between Convolutional Neural Networks and Signal Modulation Theory, *Neurocomputing* (2022), doi: <https://doi.org/10.1016/j.neucom.2022.09.088>

This is a PDF file of an article that has undergone enhancements after acceptance, such as the addition of a cover page and metadata, and formatting for readability, but it is not yet the definitive version of record. This version will undergo additional copyediting, typesetting and review before it is published in its final form, but we are providing this version to give early visibility of the article. Please note that, during the production process, errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

© 2022 Elsevier B.V. All rights reserved.



Convolutional Modulation Theory: A bridge between Convolutional Neural Networks and Signal Modulation Theory

Fuzhi Wu^{1, 2, 3, 4#}, Jiasong Wu^{1, 2, 4#}, Youyong Kong^{1, 2, 4}, Chunfeng Yang^{1, 2, 4}, Guanyu Yang^{1, 2, 4},

Huazhong Shu^{1, 2, 4*}, Guy Carrault^{2, 3, 4}, Lotfi Senhadji^{2, 3, 4}

¹Laboratory of Image Science and Technology (LIST), Key Laboratory of Computer Network and Information Integration, Southeast University, Ministry of Education, Nanjing 210096, China

²Jiangsu Provincial Joint International Research Laboratory of Medical Information Processing, Southeast University, Nanjing 210096, China

³Univ Rennes, INSERM, LTSI-UMR 1099, 35042 Rennes, France

⁴Centre de Recherche en Information Biomédicale Sino-français (CRIBs), Univ Rennes, INSERM, 35042 Rennes, France

#These authors contributed equally to this work and should be considered co-first authors

*Corresponding author Tel: 86-25-83794249, Fax: 86-25-83792698, Email: shu.list@seu.edu.cn

Abstract: Although there have been a lot of researches on convolutional neural networks (CNNs), still what happens in this black box remains a mystery. In this paper, we establish the connection between CNNs and signal modulation. From a signal modulation point of view, the forward-propagation process of CNNs can be explained as a process of modulating the input signals to the vicinity of a special energy spectrum distribution, and the back-propagation process is searching for the appropriate distribution which is better for classification or other tasks. Several experiments have been carried out to verify the modulated explanation of CNNs. Furthermore, we verify that modulating the signal to the appropriate energy spectrum distribution in advance can effectively improve the classification and segmentation accuracy.

Keywords: Convolutional neural network, signal modulation theory, energy spectrum distribution, classification, segmentation.

1. Introduction

In recent years, deep learning [1]-[10] and especially convolutional neural networks (CNNs) [11], [12] have been widely used in many research fields and industrial applications. The great success of CNNs is impressive. However, why CNN can work well is a long way from clear explanation, which has become a bottleneck restricting the development of CNNs and also their applications to areas where interpretable artificial intelligence technology is necessary as, for example, in medical domain. Therefore, researchers have proposed many interpretation frameworks, which can be roughly divided into two classes: model-specific interpretability and model-agnostic interpretability [13].

Model-specific interpretability methods can only interpret specific model types. For example, Hershey et al. [14] proposed a framework for deriving novel deep network architectures from model-based inference algorithms by unfolding the steps of the algorithm and untying the model parameters across iterations. Wu et al. [15] presented a new way to visualize, explain and understand every step of principal components analysis network (PCANet) from an energy perspective. The drawback of this practice is that these interpretation methods are not sufficiently generic, that is, if we want to use the particular type of interpretation, we have to choose the specific models and cannot use other models. Therefore, model-agnostic interpretation methods have become the focus of research in recent years. These model-free methods broadly fall into five technique types: (1) Visualization approaches try to visualize the representations to explore the pattern hidden inside a neural unit. For example, Matthew and Fergus [16] proposed deconvolution network (DeConvNet) method in which the network computations were backtracked to identify which image patches are responsible for certain neural activations. Simonyan et al. [17] demonstrated that the visualization results of image classification model by using convolutional networks (ConvNets) could be obtained by numerical optimization of input image. (2) Knowledge extraction approaches try to extract, in a comprehensible form, the knowledge acquired by a network during training and encoded as an internal representation. For example, Tan et al. [18] investigated how to use model distillation to extract complex models into transparent models. Che et al. [19] introduced a knowledge extraction method called interpretable mimic learning to learn interpretable phenotypic features, so as to make reliable predictions while imitating the performance of deep learning models. Xu et al. [20] introduced DarkSight, a visualization method used to interpret the predictions of black box classifiers on datasets inspired by the concept of dark knowledge. (3) Influence methods focus on changing the input or internal components to estimate the importance or the relevance of a feature and to better understand the network. For example, Koh and Liang [21] used influence functions to trace a model's prediction through the learning algorithm and back to its training data, thereby identifying training points that most responsible for a given prediction. Bach et al. [22] proposed the layer-wise relevance propagation algorithm to compute the relevance between decision and classifier. (4) Example-based explanation approaches try to understand the behavior of machine learning by studying particular instances of dataset. For example, Kim et al. [23] developed the maximum mean discrepancy critic (MMD-critic) which efficiently learns prototypes and criticism, designed to aid human interpretability. (5) Theoretical connection. An effective method is to establish the relationship between deep learning and some well-developed theories, and then use these theories to explain the neural networks and also guide the construction of neural networks. Some well-developed theories include: (a) Renormalization Theory. Mehta and Schwab [24] explained deep neural networks (DNNs) as a renormalization group like procedure to extract relevant features from structured data. (b) Probabilistic Theory. Patel et al. [25] developed a new probabilistic framework for deep learning based on a Bayesian generative probabilistic model. (c) Information Theory. Tishby and Zaslavsky [26] analyzed CNNs by using the theoretical framework of the information bottleneck principle. Then, Steeg and Galstyan [27] further introduced a new framework for unsupervised learning

of representations based on a novel hierarchical decomposition of information. (d) Numerical Differential Equations. Lu et al. [28] bridged deep architectures and numerical differential equations. (e) Group Theory. Paul and Venkatasubramanian [29] showed the intrinsic relations between group theory and deep networks, and explained why unsupervised deep learning works.

In signal modulation domain, many works have been conducted to optimize the traditional signal modulation system through deep learning, for example, classification of signal modulation type [30]-[34], optimization of transmitter and receiver [35], [36], quantization of L-values for gray-coded modulation [37] and spatial modulation multiple-input multiple-output (SM-MIMO) transmit antenna selection [38]. Some works establish links between the communication system and the autoencoder, interpreting the communication as an autoencoder [39]-[41]. Unlike them, we try a completely opposite problem. In this paper, we aim to provide a new model-agnostic interpretation method for CNNs by leveraging the well-studied signal modulation theory, which leads to a clear and profound understanding of CNNs, together with new insights. Specifically, we try to bridge the deep learning and signal modulation by studying the spectrum distribution of features in CNNs. We conclude that in the forward-propagation of CNNs, what happens in the black box is explained as the “generalized shifting” of energy spectrum and experiments show that different types of energy spectra will be modulated to be near a similar spectrum distribution. We conclude that the back-propagation process can be thought of as a searching process for an optimal energy spectrum distribution that is most conducive to related tasks by stochastic gradient descent methods. Experiments show that different networks will modulate features in the similar direction finally and better classified features will be closer to the optimal spectrum. Finally, the applications of our theory on one-dimensional and two-dimensional public datasets reveal that our theory is helpful for the design of CNNs. The contributions of the paper are as follows:

- 1) We propose a new interpretation framework for CNNs by using the signal modulation theory for the first time and therefore bridging the deep learning domain and signal modulation domain.
- 2) In the forward-propagation of CNNs, every operation is explained by the signal modulation theory and what happens in the black box is explained as the “generalized shifting” of energy spectrum. For the back-propagation process, it can be thought of as a searching process for an optimal energy spectrum distribution that is most conducive to classification or other tasks. Several validation experiments corroborate our theory.
- 3) We found the relation between feature spectrum distribution and task effects and several experiments reveal that our theory is helpful for the design of CNNs.

The paper is organized as follows. Signal modulation theory is reviewed in Section 2. Section 3 explains the basic operators and propagation process of CNNs. Some mysteries of CNNs are discussed in Section 4. In Section 5, two experiments are given to verify the modulation explanation of CNNs.

2. The Signal Modulation Theory

2.1 Energy spectrum shift of single-carrier amplitude modulation for the discrete-time signal

Signal modulation is an important concept in communication theory. The information carried in the low frequency modulating signal (or baseband signal) $x[n]$ will be greatly attenuated and distorted when this signal is sent on the transmission channel directly. Therefore, it is necessary to modulate $x[n]$ into a high frequency signal named modulated signal $m[n]$ that is more suitable for sending on the transmission channel. The modulated signal $m[n]$ is simply obtained by multiplying the modulating signal $x[n]$ with a high frequency carrier signal $c[n]$. Through the signal modulation method, we can use the carrier signal $c[n]$ to modulate the signal $x[n]$ so that we can get a modulated signal $m[n]$ whose characteristic parameters, such as amplitude, frequency and phase, carry the information of modulating signal $x[n]$. Sinusoidal amplitude modulation is one of the most widely used analog modulation methods in transmission of signals:

$$c[n] = \cos 2\pi f_c n, \quad (1)$$

where f_c is the frequency of $c[n]$. Then, the modulated signal $m[n]$ is given:

$$m[n] = x[n] \cdot c[n] = x[n] \cdot \cos 2\pi f_c n. \quad (2)$$

If the Fourier transforms of $x[n]$ and $m[n]$ are denoted as $X(f)$ and $M(f)$ respectively, then we obtain:

$$EM(f) = |M(f)|^2 = \frac{1}{4}|X(f + f_c) + X(f - f_c)|^2. \quad (3)$$

One of the most important applications of signal modulation is to achieve energy spectrum shift, that is, the energy spectrum of the modulating signal to be transmitted is shifted to a frequency band near a carrier signal, so that the process of sending or handling modulated signal will be more convenient than the original modulating signal.

As shown in Fig. 1, compared to the energy spectrum of the modulating signal $x[n]$, the energy spectrum of the modulated signal $m[n]$ is located around frequencies $-f_c$ and f_c , that is, signal modulation leads to energy spectrum shift and moves the energy spectrum of modulating signal towards $\pm f_c$.

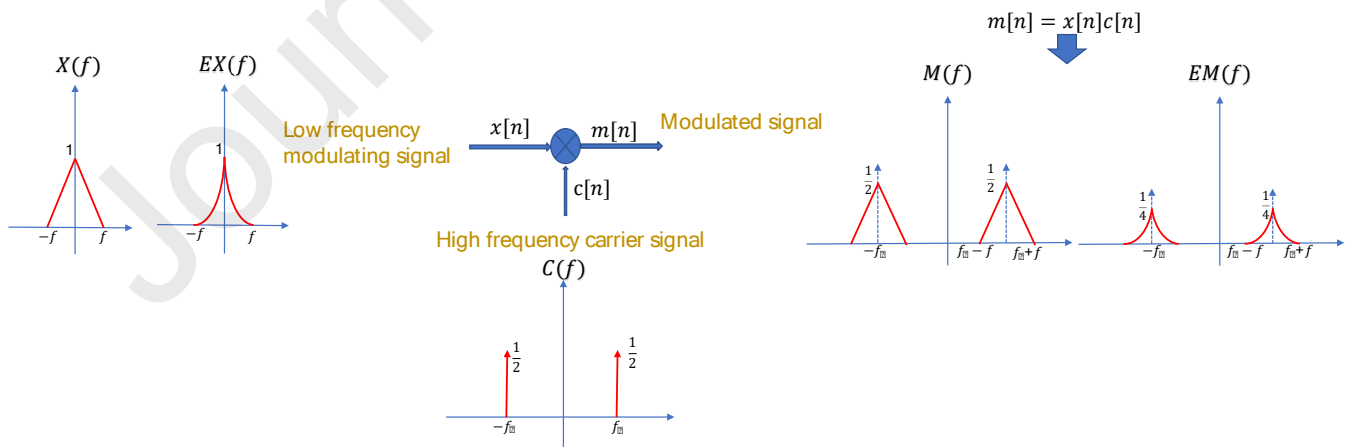


Fig. 1. The process of double sideband suppressed carrier (DSB-SC). $X(f)$ and $EX(f)$ denote spectrum and energy spectrum of modulating signal $x[n]$, respectively; $M(f)$ and $EM(f)$ denote spectrum and energy spectrum of modulated signal $m[n]$, respectively. The energy spectrum of modulating signal is moved to both sides of carrier frequency f_c .

2.2 Multi-carrier modulation and multiple-input multiple-output (MIMO) modulation

Besides single-carrier modulation shown in the above section, multi-carrier modulation is also widely used to further improve the transmission performance. As shown in Fig. 2, multi-carrier modulation divides the data stream into K_M sub-data streams $x_k[n]$, which are modulated by K_M carrier signals $\cos(2\pi f_k n)$, obtaining

$$m_k[n] = x_k[n] \cos(2\pi f_k n), k = 1, 2, \dots, K_M. \quad (4)$$

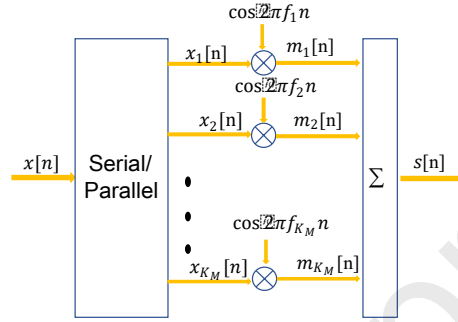


Fig. 2. General block diagram of a multi-carrier modulation. K_M is the number of carrier signals.

Then, the modulated signal is given:

$$s[n] = \sum_{k=1}^{K_M} m_k[n] = \sum_{k=1}^{K_M} x_k[n] \cos(2\pi f_k n). \quad (5)$$

Multi-carrier modulation can be implemented in many ways. A common technical approach is frequency division multiplexing (FDM), that is, the total frequency width is greater than the sum of the frequencies of each sub-channel, while ensuring that the frequency bands of the signals transmitted in each sub-channel do not interfere with each other. That is to say:

$$M_p(f) \cdot M_q(f) = 0, p \neq q, \quad (6)$$

where $M_p(f)$ and $M_q(f)$ are the spectrum of modulated signals $m_p[n]$ and $m_q[n]$ respectively. So similar to Eq. (3), the energy spectrum of Eq. (5) is given by

$$ES(f) = \left| \sum_{n=-\infty}^{+\infty} \sum_{k=1}^{K_M} x_k[n] \cos(2\pi f_k n) e^{-j2\pi f n} \right|^2 = \left| \sum_{k=1}^{K_M} M_k(f) \right|^2 = \sum_{k=1}^{K_M} EM_k(f) = \sum_{k=1}^{K_M} \frac{1}{4} |X_k(f - f_k) + X_k(f + f_k)|^2, \quad (7)$$

where $X_k(f), k = 1, 2, \dots, K_M$ are the spectrum of modulating signals $x_k[n]$; $EM_k(f)$ is the energy spectrum of $m_k[n]$. From Eq. (7), we can see that multi-carrier modulation system realizes simultaneously energy spectrum shift of multiple signals, which is shown in Fig. 3.

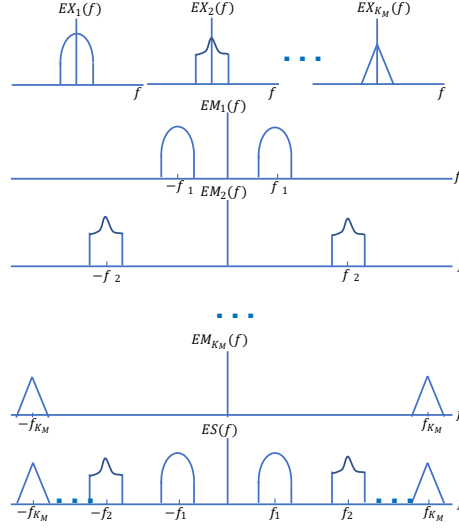


Fig. 3. The energy spectrum shift of multi-carrier modulation system.

In the actual signal transmission application, we usually combine multi-carrier technology with multiple-input multiple-output (MIMO) technology, which is a typical diversity technology used for anti-fading in communication systems. The basic principle of diversity technology is to send multiple copies carrying the same information through multiple channels. Each copy undergoes a multi-carrier modulation system on each channel. The principle block diagram of the multi-carrier MIMO system is shown in Fig. 4.

After employing spatial diversity, the input signal is divided into I_M parts on the antenna array at the transmitting end and can be expressed as:

$$\mathbf{x}[n] = [x_1[n] \quad x_2[n] \quad \cdots \quad x_{I_M}[n]]^T \in \mathbb{R}^{I_M}, \quad (8)$$

where the superscript T denotes transpose, I_M represents the number of antennas at the transmitting end. Then in each channel, $x_p[n], p = 1, 2, \dots, I_M$ is transmitted through the antenna after its corresponding multi-carrier modulation. The channel response matrix can be expressed as $\mathbf{H} \in \mathbb{R}^{O_M \times I_M}$, whose element $h_{q,p}$ represents the response coefficient from the p -th transmitting antenna to the q -th receiving antenna. Note that \mathbb{R} denotes real number domain. Then, the receiving vector after modulation of multi-carrier MIMO system can be expressed as:

$$\mathbf{r}[n] = [r_1[n] \quad r_2[n] \quad \cdots \quad r_{O_M}[n]]^T = \mathbf{H}\mathbf{s} \in \mathbb{R}^{O_M}, \quad (9)$$

where O_M represents the number of antennas at the receiving end and

$$\mathbf{s} = [s_1[n] \quad s_2[n] \quad \cdots \quad s_{I_M}[n]]^T \in \mathbb{R}^{I_M} \quad (10)$$

is the modulated signal vector.

Eq. (9) can be further refined into:

$$r_q[n] = \sum_{p=1}^{I_M} h_{q,p} s_p[n] = \sum_{p=1}^{I_M} h_{q,p} \left[\sum_{k=1}^{K_M} x_{p,k}[n] \cos(2\pi f_{p,k} n) \right], q = 1, 2, \dots, O_M, \quad (11)$$

where $x_{p,k}[n]$ corresponds to the k -th modulating signal of the p -th copy in the p -th transmit channel, $\cos(2\pi f_{p,k}n)$ corresponds to the k -th carrier signal of the p -th transmit channel.

Similar to Eqs. (3) and (7), the energy spectrum of Eq. (11) is given by

$$ER_q(f) = \sum_{p=1}^{I_M} h_{q,p}^2 ES_p(f) = \frac{1}{4} \sum_{p=1}^{I_M} h_{q,p}^2 \sum_{k=1}^{K_M} |X_{p,k}(f + f_{p,k}) + X_{p,k}(f - f_{p,k})|^2, q = 1, 2, \dots, O_M, \quad (12)$$

where $ES_p(f)$ is the energy spectrum of $s_p[n]$ and $X_{p,k}(f)$ is the frequency domain expression of signal $x_{p,k}[n]$.

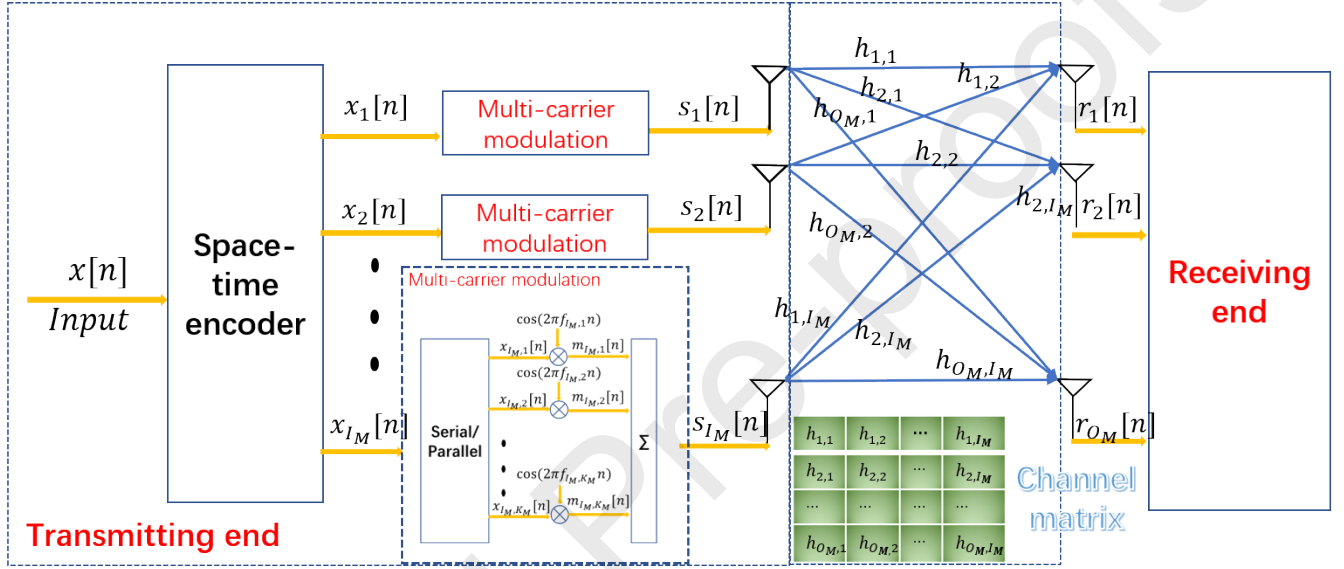


Fig.4. Signal model of multi-carrier MIMO modulation system.

3. The connection between CNNs and Modulation Theory

CNNs are hierarchical models whose inputs are raw data, such as RGB image, audio signal, and so on. Then CNNs stack various layers composed of a series of operations, such as convolution operation, pooling operation, and non-linear activation function, the purpose of which is to extract high-level semantic information from the input layer and abstract it layer by layer. This process of feature extraction is called “forward-propagation”. Finally, the last layer of the CNNs formalizes its target task (classification, segmentation, regression, etc.) into an objective function. More generally speaking, the construction of CNNs is like a process of building blocks, using the Conv-ReLU-Pooling operator as the “basic unit” in turn to “build” on the original data and “stack” layer by layer, as shown in Fig. 5. This “basic unit” will be the focus of our research. In this section, we will bridge CNNs and modulation theory and explain the correspondence in these two frames, whose main points are briefly summarized in Table 1.

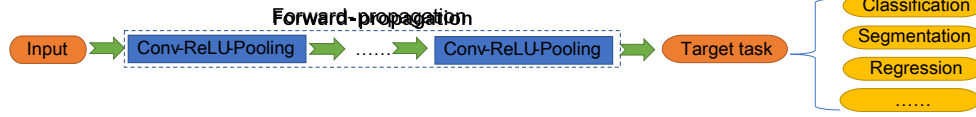


Fig. 5. The forward-propagation process of CNNs and Modulation Theory

Table 1. The connection between CNNs and modulation theory

	CNNs	Modulation Theory
Objective	Minimize the cost function. If we train with mean square error, we need to minimize $M(g; D) = \frac{1}{S_C} \sum_{i=1}^{S_C} (g(x_i) - y_i)^2$, where $D = \{(x_1, y_1), (x_2, y_2), \dots, (x_{S_C}, y_{S_C})\}$, y_i is the true label of sample x_i , g is the classifier, S_C is the number of samples. By minimizing the loss function, the appropriate parameters (filters) are obtained, that is, pretrained model, through which excellent classification or recognition effect can be obtained.	Minimize the interference of the signal in the channel and the fading of the modulating signals.
Solve	Forward-propagation Stack the Conv-ReLU-Pooling in different ways, modulate the input data to a specific energy spectrum distribution (Features $=f(\mathbf{X}; \vartheta)$, where $\mathbf{X} = \{x_1, x_2, \dots, x_{I_S}\}$ is the sample set and ϑ represents the parameters of filters), then send the result to the classifier. Back-propagation The back-propagation process is searching for the appropriate energy spectrum distribution which is best for classification, through gradient descent.	Energy spectrum shift Through different modulation methods, the energy spectrum of the original signal is moved to a higher carrier frequency, which can improve the anti-interference and anti-fading ability of the system.
Building block	Convolution Activation function (ReLU) Pooling Conv-ReLU-Pooling	Multi-carrier MIMO amplitude modulation Single-carrier amplitude modulation Multi-carrier amplitude modulation Continuous amplitude modulation

3.1 Convolution

The convolutional layer is the core layer of the CNNs, which can realize the feature extraction of the input data. The convolutional layer contains multiple convolution kernels. When the convolutional kernel works, it will regularly scan the input data, and perform matrix element multiplication and summation on the input data, thereby enhancing some features of the input data. Note that the commutativity of convolution is not very important in the practical application of CNNs, most neural network libraries regard cross-correlation function as convolution:

$$y_q[n] = \sum_{p=1}^{I_C} w_{q,p}[n] \odot x_p[n] = \sum_{p=1}^{I_C} \sum_{k=1}^{K_C} w_{q,p}[k] x_p[n+k], q = 1, 2, \dots, O_C, \quad (13)$$

where I_C is the number of input channels, O_C is the number of output channels, $x_p[n]$ denotes the p -th channel of the input feature map and $y_q[n]$ is the q -th channel of the output feature of convolution layer,

$\mathbf{w}_{q,p} = [w_{q,p}[1] \ w_{q,p}[2] \ \dots \ w_{q,p}[K_C]]^T \in \mathbb{R}^{K_C}$ is the convolutional (or correlation) kernel corresponding to the p -th input channel and the q -th output channel, and \odot is the convolution (or cross-correlation) operation. Let $I_C = O_C = 1$, then the single-channel convolution (or correlation) can be obtained:

$$y[n] = (x \odot w)[n] = \sum_{k=1}^{K_C} w[k] x[n+k], \quad (14)$$

where we discard the subscripts p and q for simplicity.

3.1.1 Connection between convolution and modulation

Let us first establish the relationship between the single-channel convolution as reported above and the multi-carrier modulation. Comparing Eq. (14) with Eq. (5), if we replace $x_k[n]$ by $x[n+k]$, $\cos(2\pi f_k n)$ by $w[k]$, in Eq. (5), and also set $K_C=K_M$, then we can obtain Eq. (14). Therefore, the single-channel convolution operation can be achieved by constructing a special multi-carrier modulation which is shown in Fig. 6. For example, if the size of the convolution kernel is $K_C=3$, the specific calculation process of convolution is shown in the left part of Fig. 6. Correspondingly, the right part is a multi-carrier modulation system with three channels. The multiplier represents scalar multiplication of vectors, the adder represents the sum of the corresponding positions of the three vectors, and then we can get the same calculation result as the convolution.

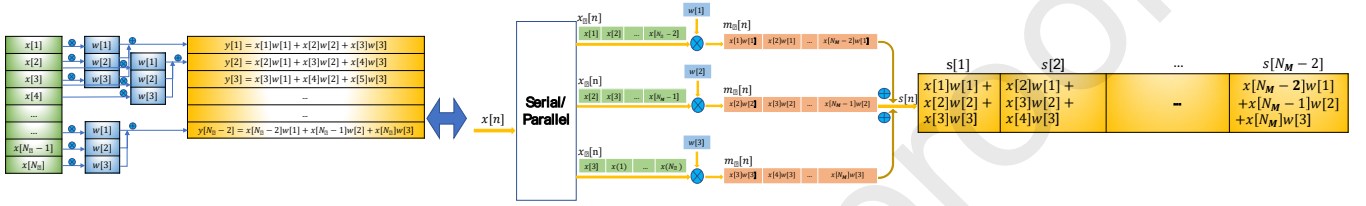


Fig. 6. The relationship between the single-channel convolution and the multi-carrier modulation. The left part is the convolution operation process in the practical application of CNNs, and the right part is the corresponding multi-carrier amplitude modulation model. N_C and N_M represent the length of input signal of convolution and modulation, respectively.

Then, we establish the relationship between the multi-channel convolution and the multi-carrier MIMO modulation, whose correspondence of parameters is shown in Table 2. Furthermore, comparing Eq. (13) with Eq. (11), if we replace $x_{p,k}[n]$ by $x_p[n+k]$, $h_{q,p}\cos(2\pi f_{p,k}n)$ by $w_{q,p}[k]$, and also set $K_C=K_M$, and $I_C=I_M$ in Eq. (11), then we can obtain Eq. (13). Therefore, the multi-channel convolution operation can be achieved by constructing a special multi-carrier MIMO modulation which is shown in Fig. 7.

For the commonly used two-dimensional image data in CNNs, we can construct a similar system, which will not be repeated here. At this point in the article, we can regard the convolution operation of CNNs as a special multi-carrier MIMO modulation system.

3.1.2 Difference between convolution and modulation

Although multi-channel convolution can be transformed into a multi-carrier MIMO modulation structure, there are still some differences between them. In signal modulation, in order to achieve a long-distance transmission of signal, we need to move the energy spectrum of low frequency signal to a position which is near the carrier energy spectrum. Generally speaking, these carrier signals are fixed and have been determined in advance. For example, we usually use the carrier signals shown in Fig. 8 in OFDM-MIMO which is one of the most widely used multi-carrier MIMO modulation methods. In contrast to the fixed carrier signals, we have no idea about what kind of energy spectrum distribution is beneficial to classification or segmentation in CNNs and also what kind of carrier signals can improve the classification accuracy of input signals. So, the ‘‘carrier parameters’’ in CNNs are usually randomly initialized, and then the optimal carrier signals are found by stochastic gradient descent method.

Table 2. Multi-channel convolution and its corresponding multi-carrier MIMO system

Multi-channel convolution	Multi-carrier MIMO modulation system
The number of data samples S_C	The number of modulating signals S_M
The number of input feature map I_C	The number of transmit antennas I_M
The number of output feature map O_C	The number of receive antennas O_M
The size of input feature map N_C	The length of modulating signal N_M
The number of convolution kernel parameters $O_C \times I_C \times K_C$	The number of subcarrier parameters $O_M \times I_M \times K_M$
The output feature map $O_C \times (N_C - K_C + 1)$	The modulated signal $O_M \times (N_M - K_M + 1)$

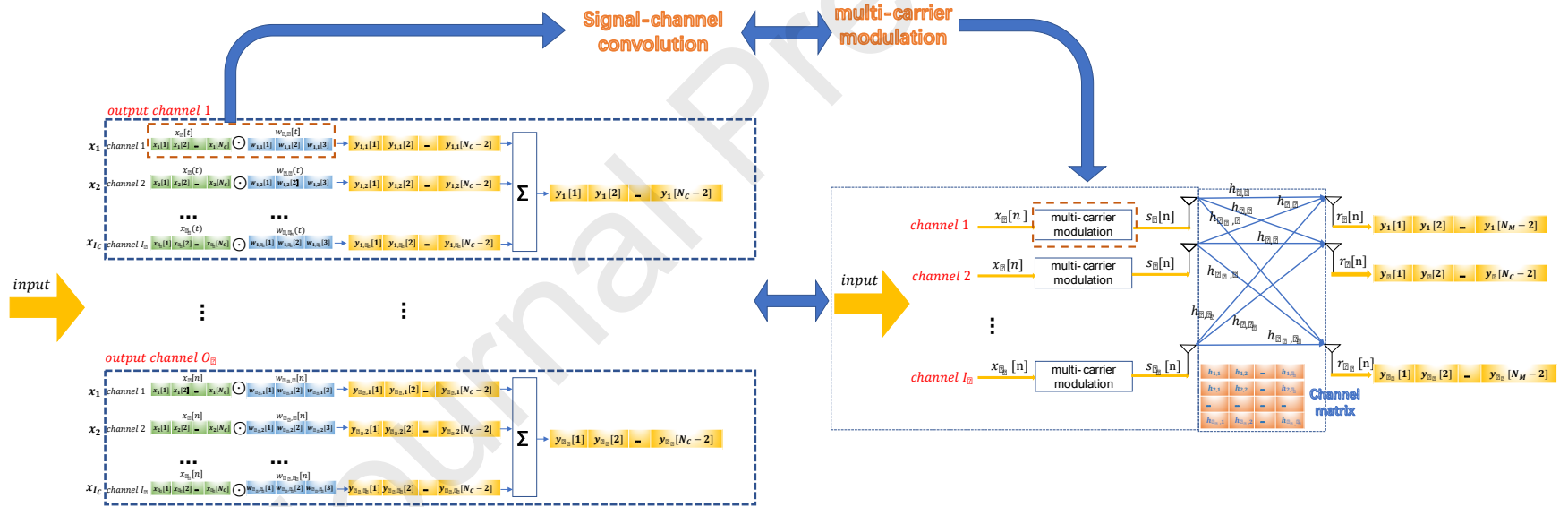


Fig. 7. The relationship between the multi-channel convolution and the multi-carrier MIMO modulation. The left part of figure shows a convolution schematic process. The input dimension is (I_C, N_C) , the convolutional kernel size is $(O_C, I_C, 3)$, and the stride is 1. The right part of figure shows the corresponding MIMO modulation system. The number of input antennas is I_M , the number of output antennas is O_M , and the number of carrier signals per channel is 3.

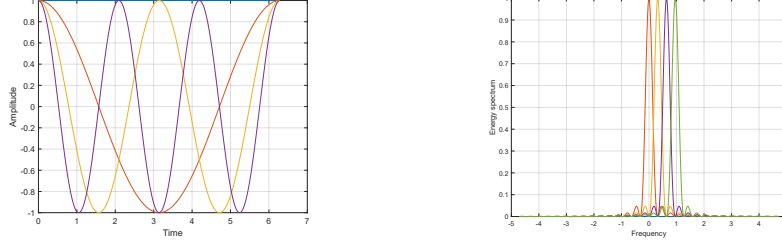


Fig. 8. The carrier signals used in the OFDM-MIMO system. The left figure is the OFDM subcarriers in time domain, and the right figure is the energy spectrum of that.

3.1.3 Energy spectrum generalized shifting of convolution operator

Since the most important role of the signal modulation system is to shift the energy spectrum of the original baseband signal into the spectrums of multi-carrier signals for a better transmission. In the following, we will check whether the convolution operator in a pretrained CNN can also realize the energy spectrum shift. In the following section, we will give the example on one-dimensional (1-D) audio classification. Example on two-dimensional(2-D) image classification is given in in the supplement document.

We analyze the signal energy spectrum shift of multi-channel convolution in a pretrained CNN for 1-D audio classification. Note that the CNN structure we used is Network-audio-1 which is shown in Fig. A1 in Appendix.

We simply choose the second convolutional layer Conv2 in Network-audio-1 with multi-channel input and multi-channel output for analysis instead of Conv1 with single-channel input and multi-channel output. The multi-channel convolution is defined in Eq. (13). The energy spectra of input and output signals are obtained by:

$$E_p^x(f) = \left| \sum_{n=-\infty}^{+\infty} x_p[n] e^{-j2\pi fn} \right|^2, p = 1, 2, \dots, I_C, \quad (15)$$

$$E_q^y(f) = \left| \sum_{n=-\infty}^{+\infty} y_q[n] e^{-j2\pi fn} \right|^2, q = 1, 2, \dots, O_C, \quad (16)$$

In order to analyze the general law, we randomly select 1000 audio segments in urbansound8K [42], which are then fed into the pretrained CNN for audio classification. The average energy spectra of the input feature map and the output feature map of Conv2 layer are respectively given by

$$\psi^x(f) = \frac{1}{S_C} \sum_{i=1}^{S_C} \sum_{p=1}^{I_C} E_{i,p}^x(f), \quad (17)$$

$$\psi^y(f) = \frac{1}{S_C O_C} \sum_{i=1}^{S_C} \sum_{q=1}^{O_C} E_{i,q}^y(f), \quad (18)$$

where S_C , I_C , and O_C are explained in Table 2, and

$$E_{i,p}^x(f) = \left| \sum_{n=-\infty}^{+\infty} x_{i,p}[n] e^{-j2\pi fn} \right|^2, \quad (19)$$

$$E_{i,q}^y(f) = \left| \sum_{n=-\infty}^{+\infty} y_{i,q}[n] e^{-j2\pi fn} \right|^2, \quad (20)$$

where $x_{i,p}[n]$ means the p -th input channel corresponding to the i -th audio sample and $y_{i,q}[n]$ means the q -th output channel corresponding to the i -th audio sample. The plots of $\psi^x(f)$ and $\psi^y(f)$ are shown in Fig. 9. Although

convolution can be transformed into a form of signal modulation, the specific carrier will be more complicated at this time. Therefore, the shift of energy spectrum of the convolutional layer is no longer a simple movement, but a process of energy spectrum shifting and reshaping (reallocation). We called the process of moving the energy spectrum to the vicinity of a specific distribution generalized shifting. From Fig. 9 we can find that the energy spectra increase in the frequency $f \in [-50, 50]$, and correspondingly the energy spectra of other parts decrease. After the Conv2 layer, the energy spectra are mainly moving to the low frequency. The concentrated area of energy spectra is narrowing during convolution. The energy spectrum generalized shifts of other convolutional layers are shown in Fig. 10.

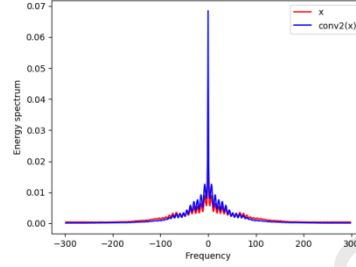


Fig. 9. The average energy spectra of $x_{i,p}[n]$ and $y_{i,q}[n]$. $\psi^x(f)$ and $\psi^y(f)$ are shown in red and blue, respectively.

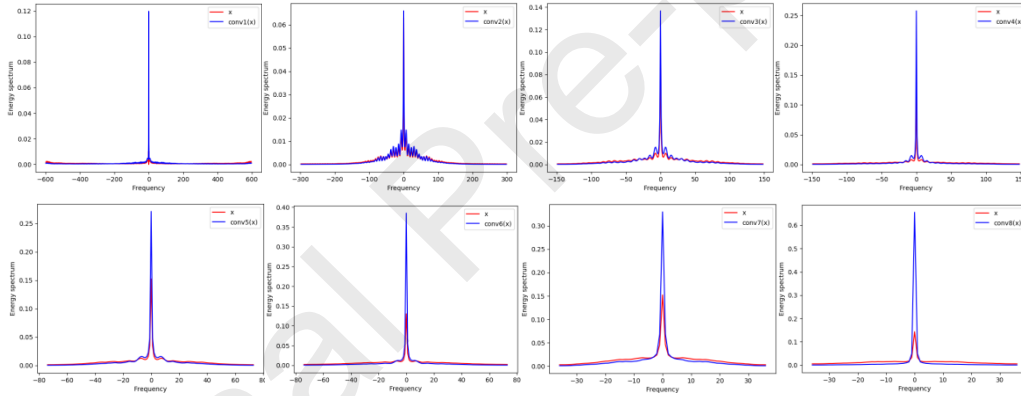


Fig. 10. The energy spectrum generalized shifts of all of the convolution layers of the Network-audio-1.

3.2 Activation function

Activation function is another important component of CNNs, and many different kinds of activation functions have been proposed, such as sigmoid, tanh, ReLU, Leaky ReLU [43], ELU [44], parametric ReLU [45], maxout [46], and sine [47], among them ReLU is probably the most popular activation function used in CNNs and the definition of ReLU is:

$$a_q[n] = \max(0, y_q[n]), \quad q = 1, 2, \dots, O_C, \quad (21)$$

where $y_q[n]$ is the q -th channel of the output feature of convolution layer defined in Eq. (13). Eq. (21) can also be expressed as:

$$a_q[n] = y_q[n]H(y_q[n]), \quad q = 1, 2, \dots, O_C \quad (22)$$

and $H(t)$ is the Heaviside function [48] given by:

$$H(t) = 1 \text{ if } t > 0 \text{ and } H(t) = 0 \text{ if } t < 0 \quad (23)$$

Comparing Eq. (22) with Eq. (2), we can see that ReLU can be constructed by a single carrier DSB-SC modulation, in which $y_q[n]$ is the modulating signal and $H(t)$ is the carrier signal.

The Fourier transform of Eq. (22) can be expressed as follows:

$$A_q(f) = \sum_{n=-\infty}^{+\infty} a_q[n] e^{-j2\pi f n}, q = 1, 2, \dots, O_c. \quad (24)$$

Note that as the support of convolution $y_q[n]$ is finite, the $A_q(f)$ is well defined.

We also check the energy spectrum generalized shift in a pretrained 1-D CNN of Network-audio-1 whose structure is shown in Fig. A1 in Appendix. We simply choose the second ReLU layer ReLU2 for the comparison with Conv2. The energy spectrum before the ReLU2 is shown in Eq. (16), and the energy spectrum after the layer ReLU2 is given by:

$$E_q^a(f) = |A_q(f)|^2, q = 1, 2, \dots, O_c. \quad (25)$$

Similarly, we randomly select 1000 audio segments in urbansound8K dataset [42], and send them to the 1-D Network-audio-1, and obtain the input and output feature maps of ReLU2 layer. Then, the average energy spectrum of each channel and each audio segment before the ReLU2 is shown in Eq. (18) and the average energy spectrum of each channel and each audio segment after the ReLU2 layer is calculated as follows:

$$\psi^a(f) = \frac{1}{S_c} \frac{1}{O_c} \sum_{i=1}^{S_c} \sum_{p=1}^{O_c} E_{i,p}^a(f), \quad (26)$$

where $E_{i,p}^a(f)$ is the energy spectrum of the q -th channel of output feature map corresponding to the i -th sample after the ReLU2 layer. O_c and S_c are explained in Table 2. The plots of $\psi^y(f)$ in Eq. (18) and $\psi^a(f)$ in Eq. (26) using ReLU are shown in Fig. 11. We can find that the energy spectra decrease in the frequency $f \in [-50, 50]$, and correspondingly the energy spectra of other parts increase. After the ReLU2 layer, the energy spectra are moving to the high frequency which is opposite to the direction of spectrum generalized shift of conv2 layer. The spectrum generalized shifts of other ReLU layers are shown in Fig. 12. Although each layer here is performing ReLU operations, the moving direction of each ReLU layer is slightly different, and we will make a more specific introduction in Section 4.

For other activation functions, we can always use Taylor's formula to convert the activation function into the following form: $a[n] = y[n]\phi[n]$ which can also be constructed by a single-carrier DSB-SC modulation. We will not give examples one by one and we verified the phenomenon of energy spectrum generalized shifts of some activation functions like sigmoid, tanh, sine, ELU on the Network-audio-1 whose activation functions are changed from ReLU to others. The plots of $\psi^y(f)$ and $\psi^a(f)$ using other activation functions instead of ReLU are shown in Fig. 13, from which we can also observe the phenomenon of energy spectrum generalized shift.

Therefore, in the perspective of signal modulation, the combination of convolution and activation function is not just a superposition of two independent linear and nonlinear operators, but a continuous process of modulation in

which energy spectrum of input will be shifted to the appropriate sidebands from two opposite directions.

Besides Conv2 and ReLU2 of Network-audio-1 for 1-D audio classification, the results of 2-D image data are shown in the supplement document.

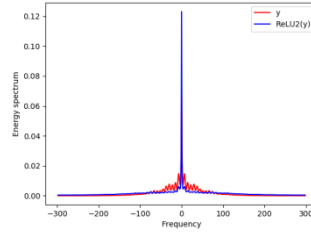


Fig. 11. The average energy spectra of $y_q[n]$ and $a_q[n]$ using ReLU function. $\psi^r(f)$ and $\psi^a(f)$ are shown in red and blue, respectively.

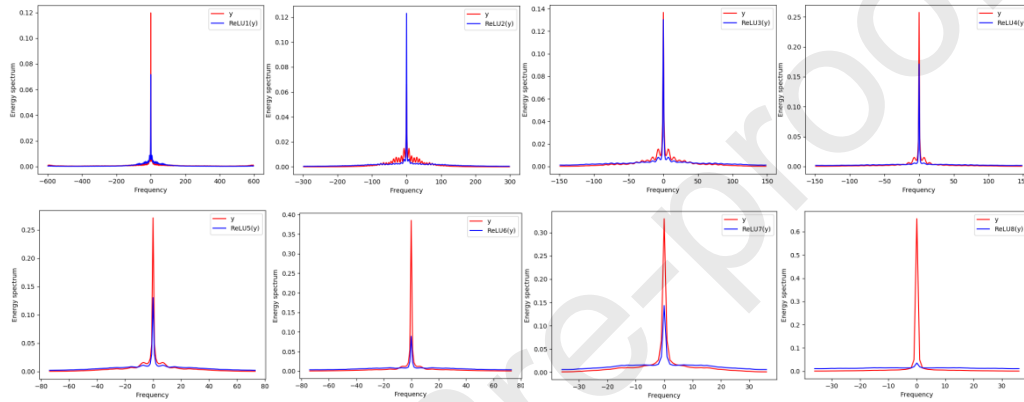


Fig. 12. The energy spectrum generalized shift of all of the ReLU layers of the Network-audio-1.

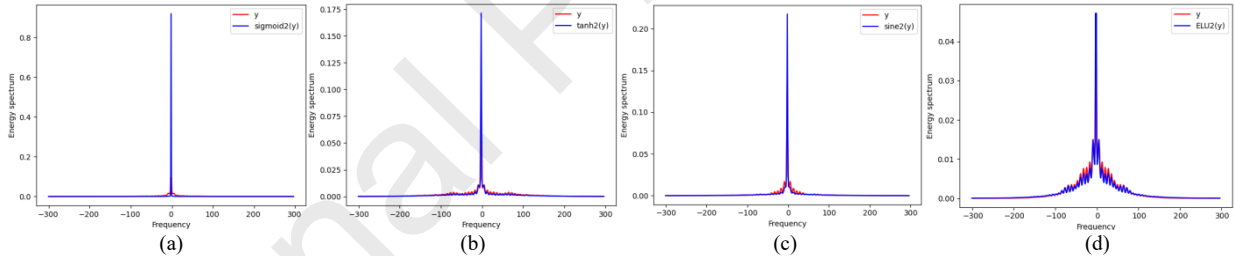


Fig. 13. The average energy spectra of $y_q[n]$ and $a_q[n]$ using other activation functions except ReLU. (a) The results of $\text{sigmoid}(y)$; (b) The results of $\text{tanh}(y)$; (c) The results of $\text{sine}(y)$; (d) The results of $\text{ELU}(y)$.

3.3 Pooling

Although Conv-ReLU can move the image energy spectrum in two different directions, and it seems that the appropriate combination of Conv-ReLU can achieve the purpose of generalized shifting the energy spectrum to the appropriate distribution. But both of them usually do not change the resolution and too high resolution will make calculations unbearable. Therefore, the introduction of pooling operations is very important in CNNs. Pooling is used in CNNs to reduce information between computational layers. Existing pooling methods mainly include pooling in spatial domain (average, max, etc) and pooling in frequency domain (Spectrum pooling) [49]. The examples of two pooling methods are shown in Fig. 14. Although one is cropped in spatial domain and the other is in frequency domain, we could not say these two kinds of methods are two completely different pooling methods. In the view of signal modulation theory, they are all modulation processes.



(a)The computational process of a typical spatial average pooling

(b)The computational process of special spectrum pooling

Fig. 14. The main pooling methods include spatial pooling and spectrum pooling.

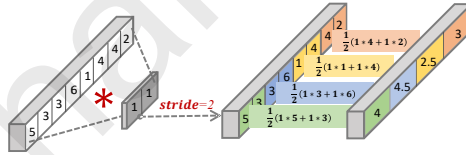
1. Spatial pooling

We try to understand a specific pooling layer, like average pooling, whose main operation is to calculate the average of the matrices in a specific range which can be easily understood as a convolution. We map the pooling function to a 1_{K_C} convolution function (step size is decided according to the resolution required by the network), as shown in Fig. 15. Therefore, this type of pooling can completely be replaced by the convolution function with specific structure as follows:

$$ap(x[n]) = \frac{1}{K_C} x[n] \odot 1_{K_C}, \quad (27)$$

where K_C is the size of convolutional kernel decided by the parameters of pooling layer. Therefore, the process of average pooling of each channel can be understood as a process of single-channel convolution. For max pooling layer, Springenberg [50] found that max pooling can simply be replaced by a convolutional layer with increased stride without loss in accuracy.

Therefore, we can regard spatial pooling as a special convolution and explain the pooling by the same way that we used to interpret the convolution, that is, we can regard spatial pooling as a way of modulation with one special carrier signal.

**Fig. 15.** The process of calculating with $input \odot 1_{K_C}$ instead of average pooling.

2. Spectral pooling

In spectral pooling, the image is truncated into a suitable size in the frequency domain. The network achieves the purpose of information compression by cropping coefficients of the low frequency of transformed feature maps. The spectral pooling can be stated as:

$$SP(x[n]) = X(f) \cdot H(f), \quad (28)$$

where $X(f)$ is the frequency representation of input feature $x[n]$, namely, $x[n]$ is firstly mapped to frequency domain by Fast Fourier transform (FFT) and then the low frequency part is shifted to the center using *fftshift*. $SP(x[n])$ is the frequency representation of output feature. $H(f)$ is a modified Heaviside function as follows

$$H(f) = \begin{cases} 1, & |f - \frac{N_p}{2}| \leq \frac{d}{2}, \\ 0, & |f - \frac{N_p}{2}| \geq \frac{d}{2}, \end{cases} \quad (29)$$

where N_p is the size of input images, d is the size of output features.

According to convolution theorem, Eq. (28) can be restated as:

$$sp(x[n]) = x[n] \odot h[n], \quad (30)$$

where $h[n]$ is the result of the inverse Fourier transform of $H(f)$. From Eq. (29), we can explain the spectral pooling as a special convolution with one specific convolution kernel, that is, spectral pooling is also a special modulation according to Section 3.1.

Through the previous description, we can consider convolution, activation, and pooling as three different modulation methods. Therefore, Conv-ReLU-Pooling, the basic operators of CNN, can be regarded as a combination of three different modulation, playing the role of the basic unit of signal modulation. Then the process of stacking Conv-ReLU-Pooling operators is actually a process of continuous signal modulation.

4. Answers to some questions in CNNs

Given the above analysis about building connection in basic units (Conv, ReLU and Pooling) of CNNs and signal modulation, we can answer the following questions from a modulation point of view:

- *Since the operator of feature extraction layer of CNNs can be interpreted as a series of signal modulation, why CNNs need modulation?*

In the signal modulation theory, we have mentioned that high frequency signals are more suitable for propagation in channels than low frequency signals, so the original low frequency modulating signal $x(t)$ needs performing modulation and shifting the energy spectrum to the position of high frequency carrier signal $c(t)$. Similarly, in the signal classification problem, if the original input signals are not suitable for classification, then, we also need to move the energy spectra of original signals to a specific frequency band and this specific energy spectrum distribution can achieve the optimal classification goal. Therefore, CNNs need to perform modulation and generalized shifting the original signal along the direction of this specific energy spectrum distribution to obtain a good classification performance. Notice that we will verify that there exists this optimal energy spectrum distribution, and different CNNs will modulate different signals to the similar energy spectrum distribution in Section 5.

- *Why each convolution in CNN is followed by ReLU function?*

In contrast to the explanation by Glorot *et al.* [51], we try to explain the question from a modulation point of view.

Through the experiments in Section 3.2, we notice that the direction of energy spectrum generalized shift of the activation layer is usually opposite to that of the convolution layer. In order to explore the more specific relationship between two energy spectrum generalized shifts, we introduce the energy spectrum difference of a layer of CNNs to represent the role of the layer in the energy spectrum generalized shift of the network. For example, the energy

spectrum differences of convolution layer and activation layer are respectively given by:

$$CDIFF(f) = \psi^y(f) - \psi^x(f), \quad (31)$$

$$ADIFF(f) = \psi^a(f) - \psi^y(f), \quad (32)$$

where $\psi^x(f)$, $\psi^y(f)$ and $\psi^a(f)$ are explained in Eqs. (17), (18) and (26), respectively. Fig. 16 shows the $CDIFF(f)$ and $ADIFF(f)$ of all of the convolution layers and ReLU layers of Network-audio-1. We can find that these two opposite directions of spectrum generalized shifts are common in the network.

Furthermore, in order to analyze the role of the two totally different kinds of structures, we no longer limit to exploring the differences of each single layer and try to accumulate the energy spectrum differences of all the convolution and activation layers respectively. The accumulation of energy spectrum differences of the convolution layer and the activation layer are respectively given by:

$$ACDIFF(f) = \sum_{n=1}^{N_{CL}} CDIFF_n(f), \quad (33)$$

$$AADIFF(f) = \sum_{n=1}^{N_{AL}} ADIFF_n(f), \quad (34)$$

where N_{CL} and N_{AL} are the number of convolution layers and activation layers of the network respectively. $CDIFF_n(f)$ and $ADIFF_n(f)$ mean the n -th $CDIFF(f)$ and the n -th $ADIFF(f)$, respectively. The plots of $ACDIFF(f)$ and $AADIFF(f)$ are shown in Fig. 17, from which we can clearly observe that the directions of the two operators (convolution and activation) are opposite and the amplitudes are close. So, in a modulation point of view, the convolution+ReLU can be seen as a delta modulation process. In other words, the energy spectrum generalized shift process of CNNs with only convolution is given:

$$FEATURE(f) = INPUT(f) + ACDIFF(f), \quad (35)$$

where $INPUT(f)$ and $FEATURE(f)$ are the energy spectrum of input and output features. The energy spectrum generalized shift process of CNNs with convolution+ReLU is given:

$$FEATURE(f) = INPUT(f) + ACDIFF(f) + AADIFF(f) = INPUT(f) + (1 - \lambda)ACDIFF(f) = INPUT(f) + \Delta ACDIFF(f), \quad (36)$$

where λ is the correlation coefficient decided by the specific convolution layer. Therefore, what CNNs with convolution+ReLU need to learn is the delta shift $\Delta ACDIFF(f)$. Convolution+ReLU is a more refined energy spectrum generalized shift than the generalized shift of single convolution. That is why CNNs with convolution+ReLU can stack more layers than CNNs only with convolution.

In addition, since the batch normalization [52] is increasingly used to optimize convolution recently, we further do an experiment to verify the role of activation functions by replacing convolution with convolution_bn (convolution+batch normalization). From Fig. 18 we can find that the reverse spectrum regulation principle of activation functions (ReLU, sigmoid, sine, tanh) is also applicable.

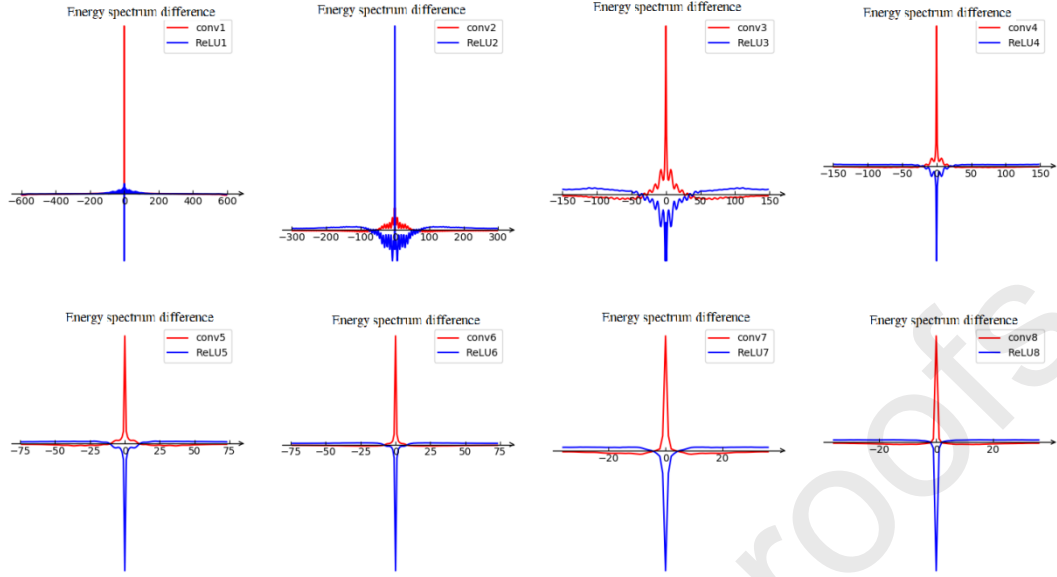


Fig. 16. The $CDIFF(f)$ and $ADIFF(f)$ of all of the convolution layers and ReLU layers of Network-audio-1.

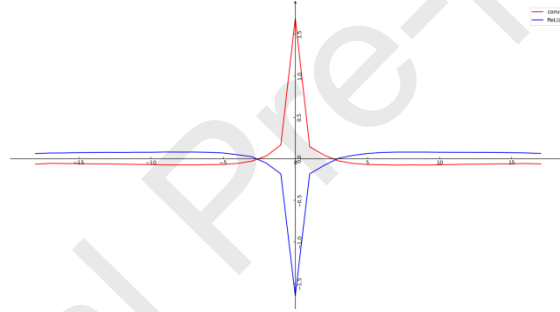


Fig. 17. The $ACDIFF(f)$ and $AADIFF(f)$ of all of the convolution layers and ReLU layers of Network-audio-1.

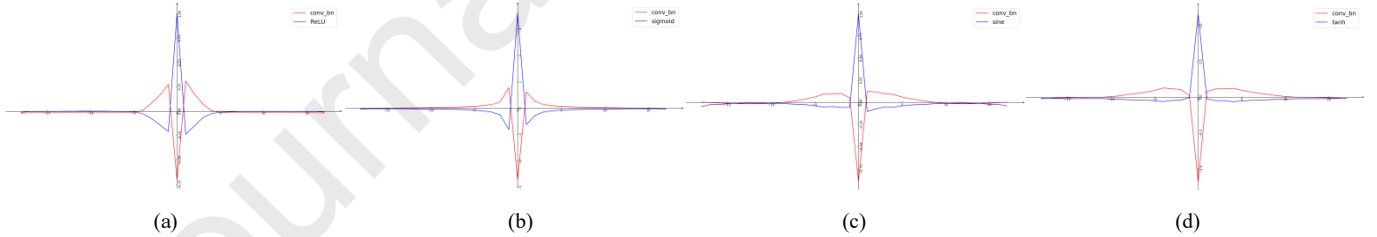


Fig. 18. The $ACDIFF(f)$ and $AADIFF(f)$ of all of the convolution layers and activation layers of Network-audio-1. Convolution layer is replaced with convolution_bn layer in (a); ReLU is replaced with sigmoid, sine and tanh in (b), (c), and (d), respectively.

- *Can another commonly used operator deconvolution be explained by signal modulation?*

Up-sampling by deconvolution [53] is a popular method for implementing semantic segmentation in deep learning. While the operation is called deconvolution, it is in fact a special kind of convolution. Deconvolution first enlarges the size of the input image to a certain ratio by interpolating 0, then rotates the convolution kernel, and finally performs convolution. Dosovitskiy et al. [54] found that Unpooling+Convolution has similar effect as deconvolution. Time domain interpolation (unpooling) does not have much effect on the resolution of the image energy spectrum, but only adds some small extra high frequency components. The main energy spectrum changes in

deconvolution are still happen in the convolution process. So, we can still regard deconvolution (a special convolution) as a special modulation.

- *Why need back-propagation in CNNs?*

In Section 3, we connect the basic units (Conv, ReLU, Pooling) of CNNs with signal modulation. We think that the forward-propagation process of CNN is a continuous modulation process, and each operator of Conv, ReLU and Pooling will modulate the input signals. Then the modulated signals will be sent to the classifier of CNNs to get the classification results. Although both signal modulation and CNNs are doing the work of modulating the energy spectrum of input signals, the biggest difference between CNNs and conventional signal modulation is the existence of back-propagation process in CNNs. In signal modulation, we know that low frequency signals are not conducive to transmit, and high frequency signals are more suitable for transmission, so what we need is just a fixed high frequency carrier which can modulate low frequency original signal into high frequency modulated signal. However, in CNNs, although it is also a continuous modulation process, we only know that the original signal cannot be easily classified, and we have no idea about what kind of energy spectrum distribution of signal is easy to classify. So, CNNs need a back-propagation process to search for a suitable energy spectrum distribution. In back-propagation, the process of adjusting parameters of convolution kernels is also the process of adjusting the energy spectrum distribution of carrier signals as shown in Fig. 19.

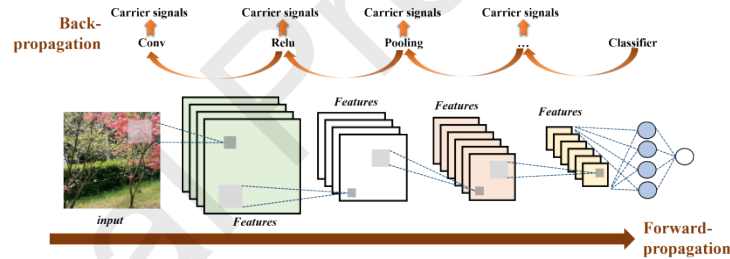


Fig. 19. The process of back-propagation is also the process of adjusting the energy spectrum of carrier signals.

5. Verification and Application Experiments

In this section, we give several experiments to verify the proposed modulation explanation theory, whose application examples are also shown. The experiments are implemented using PyTorch and Matlab on a PC machine, which sets up Ubuntu 16.04 operating system and has an Intel(R) Core(TM) i7-4790K CPU with speed of 4.00 GHz *8 and 32 GB RAM, and has one NVIDIA GTX1080-Ti GPU. Some of the models in this section are provided by torchvision in PyTorch, and some hyperparameters of models have been modified to meet the video memory requirements.

5.1 Verification experiment

In the proposed modulation theory, CNNs are continuous energy spectrum modulation process. The verification experiments are divided into three parts:

- (1) In the forward-propagation process, what the network does is moving the energy spectrum distribution of input signals. Similar to the process of signal modulation, that the carrier signal modulates different modulating signals

to the vicinity of energy spectrum of the carrier signal, the CNNs (a continuous spectrum modulation process) modulate different input signals to the similar energy spectrum distribution. Experiment will be conducted to verify whether input signals with different energy spectrum distributions will be moved to the similar distribution.

- (2) During the back-propagation process, the CNNs continuously adjust the Conv-ReLU-Pooling (carrier signal parameters) to make the energy spectrum of the modulated signal change in a specific direction. In the second part, we will verify that the back-propagation process of networks of different structures (composed of basic Conv-ReLU-Pooling structure) essentially moves the energy spectrum along different routes toward the similar direction of energy spectrum distribution which is beneficial for classification.
- (3) Since CNNs actually act on each individual image, we will explore how different the output spectra are if two different images are taken from one category and find the association between optimal spectrum distribution and individual classification effects.

5.1.1 Forward-propagation

In this part, the network we choose is a 152-layer residual network (Since we study a generalized modulation process with forward propagation, here we test on an untrained network with random initialization). Ten different categories of ImageNet [55] are randomly selected and the specific categories are shown in the Table 4.

Then we explore the average energy spectrum distributions of the input images and the output features of CNNs, as shown in Fig. 20. The average energy spectra of input images and output features of each category are given:

$$AE^{input}(\alpha, \beta) = \frac{1}{N_S} \sum_{i=1}^{N_S} E_i^{input}(\alpha, \beta), \quad (37)$$

$$AE^{feature}(\alpha, \beta) = \frac{1}{N_S} \sum_{i=1}^{N_S} E_i^{feature}(\alpha, \beta), \quad (38)$$

where N_S is the number of samples of each category, and

$$E_i^{input}(\alpha, \beta) = \frac{1}{C_I} \sum_{p=1}^{C_I} E_{i,p}^{input}(\alpha, \beta) = \frac{1}{C_I} \sum_{p=1}^{C_I} \left| \sum_{u=1}^{N_{I_1}} \sum_{v=1}^{N_{I_2}} input_{i,p}(u, v) e^{-j2\pi \left(\frac{\alpha u}{N_{I_1}} + \frac{\beta v}{N_{I_2}} \right)} \right|^2, \quad (39)$$

$$E_i^{feature}(\alpha, \beta) = \frac{1}{C_F} \sum_{q=1}^{C_F} E_{i,q}^{feature}(\alpha, \beta) = \frac{1}{C_F} \sum_{q=1}^{C_F} \left| \sum_{u=1}^{N_{F_1}} \sum_{v=1}^{N_{F_2}} feature_{i,q}(u, v) e^{-j2\pi \left(\frac{\alpha u}{N_{F_1}} + \frac{\beta v}{N_{F_2}} \right)} \right|^2, \quad (40)$$

where C_I and C_F are the number of channels of input images and output features. $input_{i,p}(u, v)$ is the element of $\mathbf{input}_{i,p} \in \mathbb{R}^{N_{I_1} \times N_{I_2}}$, which is the p -th channel of the i -th sample of input images. $feature_{i,q}(u, v)$ is the element of $\mathbf{feature}_{i,q} \in \mathbb{R}^{N_{F_1} \times N_{F_2}}$, which is the q -th channel of the i -th sample of output features. For each category, we calculate Eqs. (39) and (40) after averaging 1300 samples in ImageNet with *Natural Image Statistical Toolbox* [56]. The curves in the image represent 60%, 80%, and 90% of the image energy spectrum for each section

from the center to the periphery. Algorithm 1 provides the pseudo-code for how we get the average energy spectrum for each section. Notice that the 80% and 90% curves of the modulated image are so close that they may be indistinguishable but the curves of input image are not so. From Table 4, we can find that CNN normalizes different categories of images with different initial energy spectrum distributions to the similar energy spectrum distribution during the process of forward-propagation.

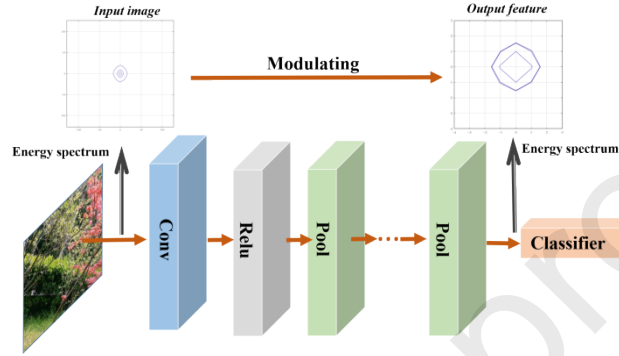


Fig. 20. The energy spectrum distributions of the input images and the output features of CNNs. The lower half is the forward-propagation of CNNs. The input image is regarded as the modulating signal, and the output feature as the modulated signal. In the upper part the energy spectrum distributions of the modulating signal and the modulated signal are compared.

Algorithm 1: Average energy spectrum

Input: $N_s, C, images(input_i)=input_1, input_2, \dots, input_{N_s}$

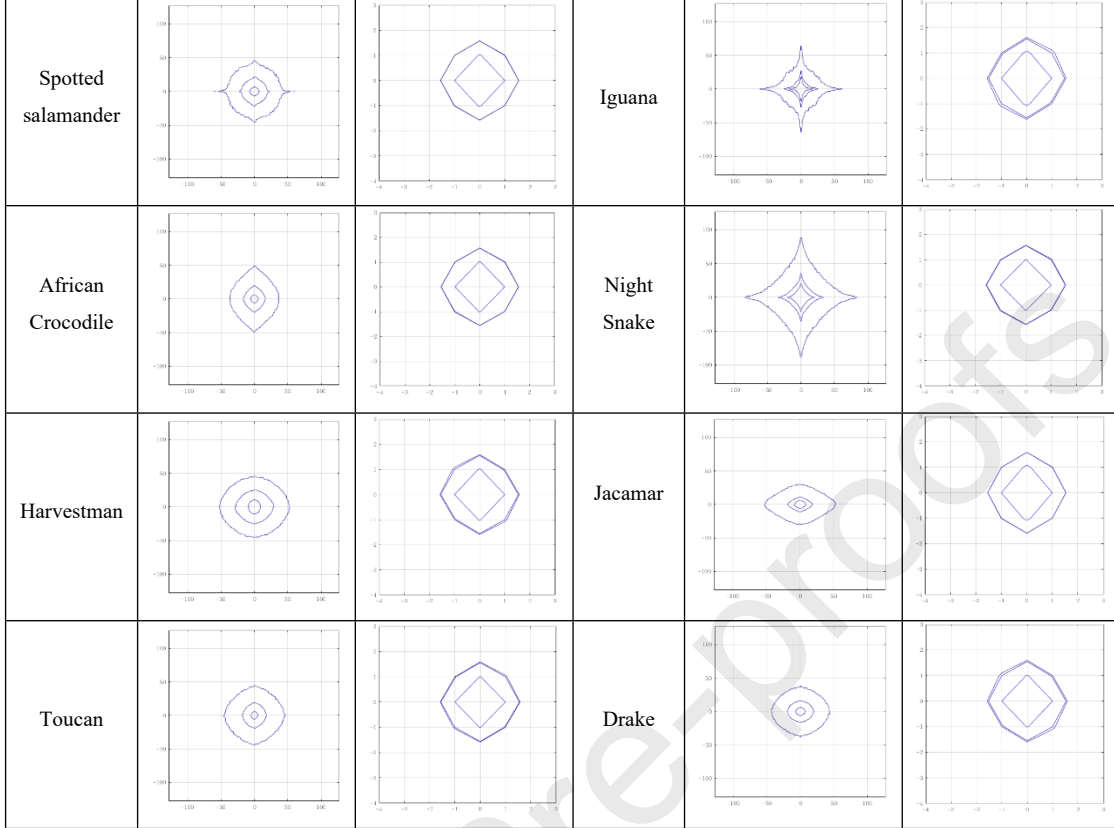
```

1 for  $i = 1; i \leq N_s$  do
2   for  $p = 1; p \leq C$  do
3      $fftImg = |fft2(input_{i,p})|^2$ ;
4      $fftImg = \frac{fftImg}{sum(fftImg)}$ ;
5      $averageFFT = averageFFT + fftImg$ ;
6  $averageFFT = \frac{averageFFT}{N_s * C}$ ;
7  $averageFFT = fftshift(averageFFT)$ ;
8  $[thresholds, ndx] = sort(-averageFFT(:)); thresholds = -thresholds$ ;
9  $energySections = cumsum(thresholds)$ ;
10  $ndx90 = \min(abs(energySections - 0.9)); th90 = thresholds(ndx90)$ ;
11  $ndx80 = \min(abs(energySections - 0.8)); th80 = thresholds(ndx80)$ ;
12  $ndx60 = \min(abs(energySections - 0.6)); th60 = thresholds(ndx60)$ ;

```

Table 4. The average energy spectrum distribution of ten different categories of images. The first and the fourth columns are category names. Columns 2 and 5 are the average energy spectra of images (modulating signals) of each category. Columns 3 and 6 are the average energy spectra of the features extracted by ResNet152.

category	Input (modulating signal)	Output (modulated signal)	category	Input (modulating signal)	Output (modulated signal)
Tench			Goldfish		



5.1.2 Back-propagation

In this part, the back-propagation process of several different CNNs, including Alexnet_bn, VGG19_bn, Resnet18, Resnet152, and Resnext50, will be explored. Similar to Section 5.1.1, we randomly select ten different classes in ImageNet to complete training. In order to intuitively display the changing process of the average energy spectrum distribution of images, we use 3-D coordinates $(\mu, \nu, epoch)$ to represent the energy spectrum changes of the modulated signal of each epoch.

One of the most famous scale invariances of natural image statistics [57], [58] is $1/\rho$ law, which states that the amplitude of the averaged Fourier spectrum $AF(\rho)$ of the ensemble of natural images obeys a distribution [59]:

$$AF(\rho) \propto 1/\rho, \quad (41)$$

where ρ is the radius from the spectrum center. Therefore, the average energy spectrum $AE(\rho)$ obeys the distribution:

$$AE(\rho) \propto 1/\rho^2, \quad (42)$$

which means that the average energy spectrum of the natural images, after averaging over orientations, lies approximately on a straight line on a log-log scale. Therefore, the specific relationship of $\log(AE(\rho))$ and

$\log(1/\rho^2)$ of five different CNNs can be obtained by linear regression, that is, $\log(AE(\rho)) = \mu \log(1/\rho^2) + \nu$.

In this experiment, 200 epochs are conducted in the process of back-propagation. For each epoch, (μ, ν) are calculated and the plots are shown in Fig. 21. We can find that although the paths are not exactly the same, the back-propagation processes of five different CNNs are similar: the networks move the image energy spectrum towards the similar end, which is beneficial to classification. Notice that the starting points of five different CNNs are totally different mainly due to the different modulation results caused by different CNNs with totally different structures.

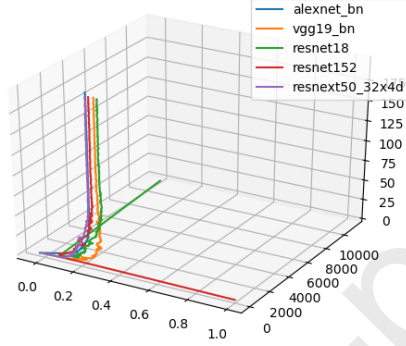


Fig. 21. The energy spectrum modulation process of five different CNNs (AlexNet_bn, VGG19_bn, Resnet18, Resnet152, and Resnext50). The abscissa, ordinate and vertical coordinates in the figure represent μ , ν , and *epoch*, respectively.

5.1.3 Connections between spectrum distribution and task

In Section 5.1.1 we verified that CNNs modulate different types of image spectra to a similar spectrum distribution, and in Section 5.1.2, we introduced that in backpropagation, different CNNs will follow different paths to search the best spectrum distribution in the similar direction. The task of CNNs is to accomplish the set goal for each image. So, if we want to better understand CNNs with modulation theory, we need to relate the spectrum of a specific image to the target. We will then study the connection between the optimal spectrum distribution and the effect of the task.

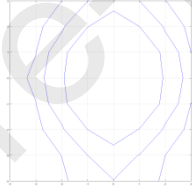
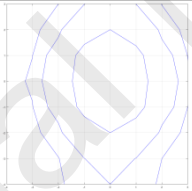
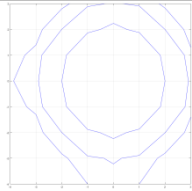
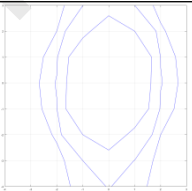
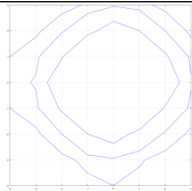
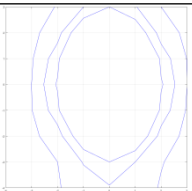
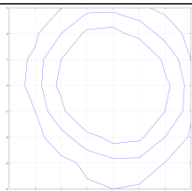
The network we choose is the 152-layer residual network (Since we study correlation between the classification results and spectrum, here we test on a trained network). We choose the specific category of “Tench” in ImageNet. In order to compare the association between spectrum and classification results, we select 20 images with the best classification results and 20 images with the worst classification results. Notice that we consider the images with a high probability of Top-1 being “Tench” in the classification results as good classifications and vice versa as bad classifications.

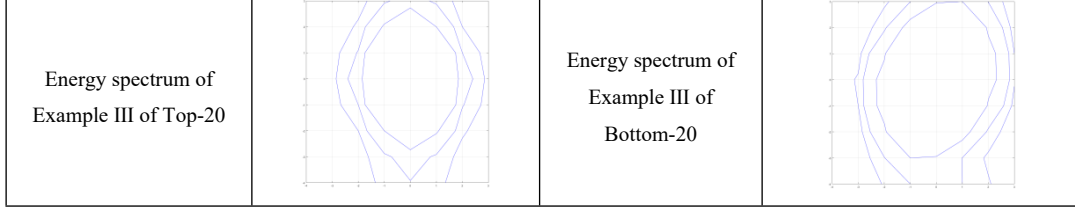
We take the same study approach as in Section 5.1.1, but here we no longer do a study of the average spectrum of a category, but specifically compare each individual image (Eq. (40)). We compare the optimal spectrum distribution, the average spectrum distribution of the Top-20, and the spectrum distribution of the individual image in the Top-20. We compare the optimal spectrum distribution, the average spectrum distribution of the Bottom-20, and the spectrum distribution of the individual image in the Bottom-20. Table 5 shows the results. Top-20 indicates the

20 best-performing images, and Bottom-20 indicates the 20 worst-performing images. We can find that for each individual image, good classified images will be closer to this optimal spectrum distribution, and the less well classified ones will be further away.

At last, combined with our validation experiments, we give our final understanding of the CNNs as a tool to finish modulating. The CNNs are trying to do a spectrum normalization process to modulate different types of images to a similar spectrum distribution nearby, and by gradient descent methods the backpropagation process searches for an optimal spectrum distribution which is beneficial to the set task and better performed cases will be more inclined to be closer to. We have given our understanding of deep learning as a black box, how can we use modulation theory to improve deep learning? Since images with spectrum closer to the optimal spectral distribution tend to perform better for the target task, having an energy spectrum generalized shift process in advance become a natural fit in cases where we want to exploit the spectrum for the set tasks. We give some application experiments on 1-D audio classification task and 2-D image segmentation in the next section.

Table 5. Comparison between spectrum distributions of individual image in one certain class. The second column are 20 images with the best classification results. The fourth column are 20 images with the worst classification results. Top-20 indicates the 20 best-performing images, and Bottom-20 indicates the 20 worst-performing images.

Optimal spectrum distribution			
Average energy spectrum distribution of Top-20		Average energy spectrum distribution of Bottom-20	
Energy spectrum of Example I of Top-20		Energy spectrum of Example I of Bottom-20	
Energy spectrum of Example II of Top-20		Energy spectrum of Example II of Bottom-20	



5.2 Application experiment

In this section, we will show the application examples of the proposed modulation explanation theory. Since we have explained that the CNNs are essentially a continuous signal modulation process, what the back-propagation of the CNNs really does is to find an energy spectrum distribution which is most conducive to classification or other tasks, and input signals with different energy spectrum distributions will be moved to this optimal spectrum distribution in the forward-propagation. Naturally, we consider whether this energy spectrum distribution can be used in some practical applications to guide the network design and to obtain an improved result in some tasks, for example, classification, segmentation, etc.

Here the energy spectrum generalized shift, that is, energy spectrum difference of CNN is given:

$$DIFF(f) = FEATURE(f) - INPUT(f), \quad (43)$$

where $INPUT(f)$ and $FEATURE(f)$ are the energy spectrum of input and output features. As shown in Fig.22, we firstly train a CNN to obtain the $FEATURE(f)$ and $INPUT(f)$. Then we can calculate the $DIFF(f)$ by linear fitting of Eq. (43). Finally, we try and see if $DIFF(f)$ can be used to improve the performance of CNNs. In the following two experiments, we shift the input spectrum to a better spectrum distribution in advance by determined carrier signals. We called the process post-shift. We will verify whether post-shift signals will get better results compared to raw data.

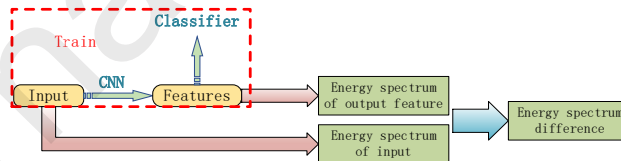


Fig. 22. Energy spectrum difference of CNN can be obtained from the energy spectrum of input and output feature of a trained CNN.

5.2.1 Audio classification

In this experiment, we try to move the energy spectrum of original audios to the vicinity of the energy spectrum distribution which is conducive to classification in advance, and then to see if the energy spectrum post-shift process can obtain an improved result in audio classification as shown in Fig. 23.

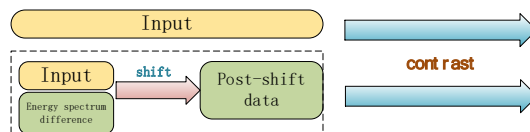


Fig. 23. Original audios and post-shift audios are sent to the same network to compare the classification results.

We first take two datasets Urbansound8k [42] and Dcase2016 [60] as Data-audio-A and Data-audio-B respectively. Data-audio-A is used to get the energy spectrum difference in advance and Data-audio-B is the part we use to validate the results. Then we create two simple CNNs: Network-audio-2 and Network-audio-3 whose detailed convolution and pooling hyperparameters are shown in Fig. A2 in Appendix. The stochastic gradient descent algorithm and cross-entropy loss are adopted in both CNNs, and each CNN is trained with 200 epochs. As shown in Fig. 24, experiment is carried according to the following three steps and Algorithm 2 shows the pseudo-code:

- (1) Data preprocessing. We firstly extract six features from these audios, that is, Short Time Fourier Transform (STFT), Mel Frequency Cepstrum Coefficient (MFCC), Chroma, Mel Spectrogram, Spectral Contrast and Tonnetz feature, and then combine these features as input.
- (2) Acquisition of energy spectrum difference. We firstly train a 5-layer classification network (Network-audio-2) with dataset Data-audio-A. Then making use of the energy spectrum distribution of features obtained by the trained classification network, we can obtain the $DIFF(f)$ by linear regression method with the energy spectrum $INPUT(f)$ and $FEATURE(f)$.
- (3) Comparison results. We firstly train a 3-layer classification network (Network-audio-3) with original Data-audio-B. Then we shift the energy spectrum of Data-audio-B with the obtained energy spectrum difference, and then send the post-shift audios to the same network but without pooling layer to compare original results. As shown in Fig. 25, we can find that the classification accuracy is effectively improved by shifting the energy spectrum of signal to the vicinity of a certain excellent energy spectrum distribution in advance.

Algorithm 2: Spectrum post-shift algorithm

```

Input: Data_A, Data_B
1 for  $i = 1; i \leq epochs$  do
2    $model\_complex.train(Data\_A);$ 
3  $feature\_A=model\_complex.Encoder(Data\_A);$ 
4  $model\_regression.train(fft2(Data\_A),fft2(feature\_A));$ 
5 for  $i = 1; i \leq epochs$  do
6    $transformed\_data=model\_regression.val(fft2(Data\_B));$ 
7    $transformed\_data=ifft(transformed\_data);$ 
8    $model\_simple.train(transformed\_data);$ 

```

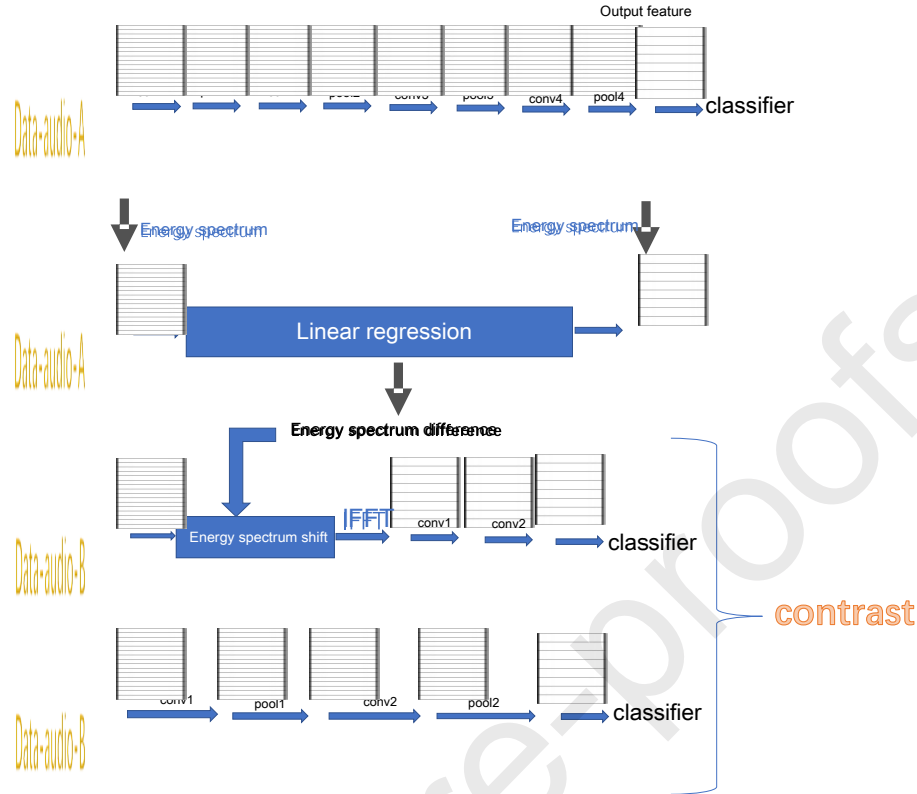


Fig. 24. The flow chart of 1-D audio classification experiment. An excellent energy spectrum distribution is obtained through a CNN, then the energy spectrum difference is used to shift the energy spectrum of signal to this good distribution in advance, and finally the classification results of the post-shift signal and the original signal are compared.

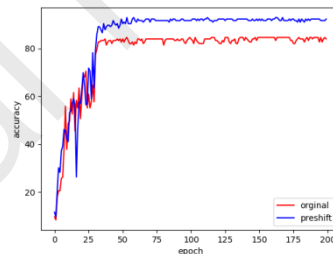


Fig. 25. Comparison results of the classification accuracy of the original signal and post-shift signal. The blue one is the accuracy curve of post-shift audio in Data-audio-B, and the red one is that of the original audio.

5.2.2 Image segmentation

In this experiment, we try to move the energy spectrum of original images to the vicinity of the energy spectrum distribution which is conducive to segmentation in advance, and then to see if energy spectrum post-shift images are more conducive to guiding semantic segmentation than the original images.

We firstly divide the Pascal VOC2012 [61] dataset into two parts: Data-image-A and Data-image-B, and each containing half of the original images. Data-image-A is used to get the energy spectrum difference in advance and Data-image-B is the part we use to validate the results. In this experiment, we construct two segmentation convolution networks: Network-image-4 and Nwtwork-image-5 whose detailed convolution, deconvolution and pooling

hyperparameters are shown in Fig. A3 in Appendix. The stochastic gradient descent algorithm and cross-entropy loss are adopted in both CNNs, and each CNN is trained with 200 epochs. Then as shown in Fig. 26, experiment is carried according to the following steps:

- (1) Acquisition of energy spectrum difference. We firstly train Network-image-4 with Data-image-A. Then making use of the energy spectrum distribution of features obtained by the trained segmentation network, we can obtain the $DIFF(f)$ by linear regression method with the energy spectrum $INPUT(f)$ and $FEATURE(f)$.
- (2) Comparison results. We firstly train Network-image-5 with dataset Data-image-B to get the original segmentation accuracy. Then we replace the skip connection in Network-image-5 with energy spectrum generalized shift by the obtained energy spectrum difference, that is, we use post-shift images instead of original images guiding the segmentation layer. The result is shown in Table 6, and we can find that moving energy spectrum distribution of the original image to a certain optimal energy spectrum distribution in advance can achieve better results on mean accuracy and mean IU for image segmentation task.

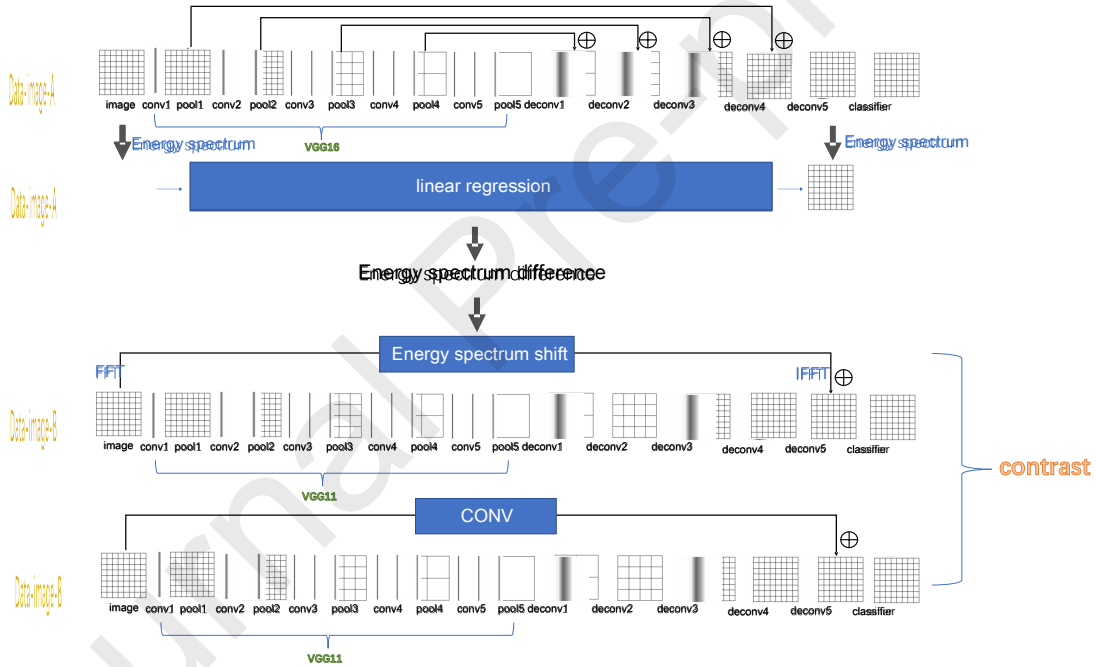


Fig. 26. The flow chart of 2-D image segmentation experiment. An optimal carrier signal through a CNN is obtained first, and then this carrier signal is used to modulate the image signals to improve the segmentation performance.

Table 6. Comparison of Mean accuracy and Mean IU verification accuracy of the original image signal and modulated image signal in Data-image-B.

	Mean accuracy	Mean IU
Original images	0.4737	0.3621
Post-shift images	0.5169	0.4017

6. Conclusion and discussion

In this paper, we explained CNNs from a new perspective based on signal modulation theory. In our explained

framework, each basic operator in CNNs can be seen as the modulation of the signals. The combination of convolution and ReLU layer is thought of as a special delta modulation. The stack of basic module of Conv-ReLU-Pooling in CNNs is essentially a continuous modulation of the signal energy spectrum. What happens in forward-propagation is regarded as moving the energy spectrum of different input signals to the certain energy spectrum distribution. For the back-propagation process, it can be thought of as a searching process for an optimal energy spectrum distribution that is most conducive to classification or other tasks and the optimal carrier signals will be found by gradient descent during the process. Our experiments prove that the CNNs modulate the original signals with different energy spectrum distributions to the similar spectrum distribution to achieve better classification goal. CNNs with different structures will move the energy spectrum distribution along different paths but towards the similar end and images that are shifted closer to the optimal spectrum distribution will tend to perform better. Furthermore, shifting the input signal to the vicinity near the optimal energy spectrum distribution in advance can significantly improve the performance of classification and the shifted images can guide the semantic segmentation more effectively.

In this work, we explain what happens in CNNs by studying the spectrum distribution of features in the spatial domain. However, in most classical CNNs, the spatial domain of the features is continuously down-sampled and the dimensionality channel domain of the features or the width of CNN is constantly increasing. Attempts to better understand CNNs in the channel domain can be of great interest to the study of interpretability in deep learning. The next phase of our research will focus on the channel domain in more depth. We see this will be an important direction for further understanding of CNNs and needs further study.

Appendix: Network detail

We list all the networks we used throughout this article in detail. The structure of Network-audio-1 is shown in Fig. A1.

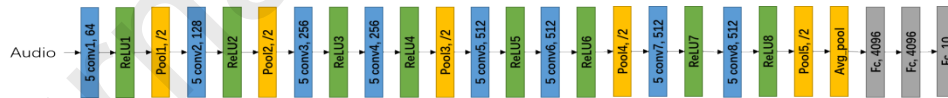


Fig. A1. The structure of Network-audio-1. Note that this CNN is modified appropriately from VGG-11 to process 1-D audio data.

The structures of Network-audio-2 and network-audio-3 are shown in Fig. A2.

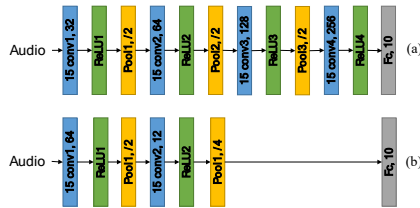


Fig. A2. The structures of Network-audio-2 and Network-audio-3 used in Section 5.2.1. (a) Network-audio-2 is used to obtain the optimal energy spectrum difference; (b) The obtained optimal energy spectrum difference is used to optimize the Network-audio-3.

The structures of Network-image-4 and Network-image-5 are shown in Fig. A3.

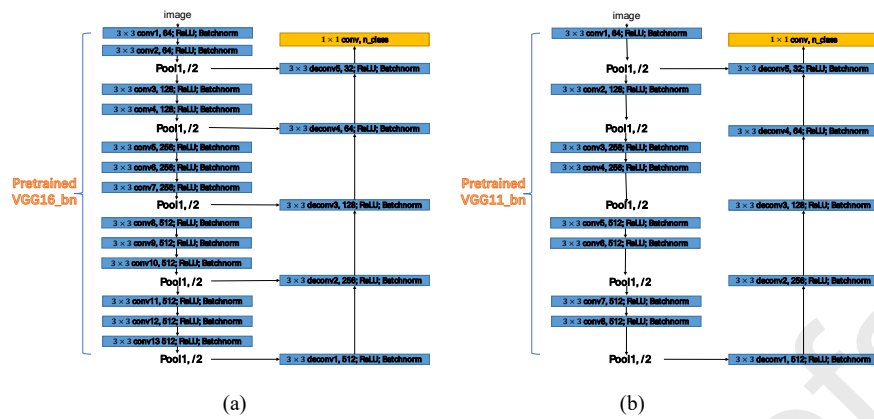


Fig. A3. The structures of Network-image-4 and Network-image-5 used in Section 5.2.2. (a) Network-image-4 is used to obtain the optimal energy spectrum difference; (b) The obtained optimal energy spectrum difference is used to optimize the Network-image-5.

ACKNOWLEDGEMENT

This work was supported in part by the National Key Research and Development Program of China (No. 2021ZD0113202), and in part by the National Natural Science Foundation of China under Grants 61876037, 62171125, 31800825, 61871117, 61871124, 61773117, 61872079, and in part by the Pre-Research Foundation of 50912040302, and in part by INSERM under the Grant the calls IAL and IRP. The authors would like to thank the Big Data Computing Center of Southeast University for providing the facility support on the numerical calculations in this paper.

References

- [1] Hinton G E, Osindero S, Teh Y W. A fast learning algorithm for deep belief nets[J]. *Neural computation*, 2006, 18(7): 1527-1554.
- [2] Hinton, G, E, et al. Reducing the Dimensionality of Data with Neural Networks.[J]. *Science*, 2006.
- [3] Bengio Y , Lamblin P , Popovici D , et al. Greedy layer-wise training of deep networks[C]// *Advances in Neural Information Processing Systems 19, Proceedings of the Twentieth Annual Conference on Neural Information Processing Systems*, Vancouver, British Columbia, Canada, December 4-7, 2006. DBLP, 2007.
- [4] Lee H , Grosse R , Ranganath R , et al. Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations[C]// *Proceedings of the 26th Annual International Conference on Machine Learning, ICML 2009*, Montreal, Quebec, Canada, June 14-18, 2009. ACM, 2009.
- [5] Vincent P, Larochelle H, Bengio Y, et al. Extracting and composing robust features with denoising autoencoders[C]// *Proceedings of the 25th international conference on Machine learning*. 2008: 1096-1103.
- [6] LeCun Y, Bengio Y, Hinton G. Deep learning[J]. *nature*, 2015, 521(7553): 436-444.
- [7] Bengio Y. Learning deep architectures for AI[M]. Now Publishers Inc, 2009.
- [8] Deng L, Yu D. Deep learning: methods and applications[J]. *Foundations and trends in signal processing*, 2014, 7(3-4): 197-387.
- [9] Bengio Y, Courville A, Vincent P. Representation learning: A review and new perspectives[J]. *IEEE transactions on pattern analysis and machine intelligence*, 2013, 35(8): 1798-1828.
- [10] Schmidhuber J. Deep learning in neural networks: An overview[J]. *Neural networks*, 2015, 61: 85-117.
- [11] LeCun Y, Boser B, Denker J S, et al. Backpropagation applied to handwritten zip code recognition[J]. *Neural computation*, 1989, 1(4): 541-551.
- [12] LeCun Y, Bottou L, Bengio Y, et al. Gradient-based learning applied to document recognition[J]. *Proceedings of the IEEE*, 1998, 86(11): 2278-2324.
- [13] Adadi A, Berrada M. Peeking inside the black-box: A survey on Explainable Artificial Intelligence (XAI)[J]. *IEEE Access*, 2018, 6: 52138-52160.
- [14] Hershey J R, Roux J L, Weninger F. Deep unfolding: Model-based inspiration of novel deep architectures[J]. *arXiv preprint arXiv:1409.2574*, 2014.
- [15] Wu J, Qiu S, Kong Y, et al. PCANet: An energy perspective[J]. *Neurocomputing*, 2018, 313: 271-287.
- [16] Zeiler M D, Fergus R. Visualizing and understanding convolutional networks[C]// *European conference on computer vision*. Springer, Cham, 2014: 818-833.
- [17] Simonyan K, Vedaldi A, Zisserman A. Deep inside convolutional networks: Visualising image classification models and saliency maps[J]. *arXiv preprint arXiv:1312.6034*, 2013.
- [18] Tan S, Caruana R, Hooker G, et al. Distill-and-compare: Auditing black-box models using transparent model distillation[C]// *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*. 2018: 303-310.
- [19] Che Z, Purushotham S, Khemani R, et al. Distilling knowledge from deep networks with applications to healthcare domain[J]. *arXiv preprint arXiv:1512.03542*, 2015.
- [20] Xu K, Park D H, Yi C, et al. Interpreting deep classifier by visual distillation of dark knowledge[J]. *arXiv preprint arXiv:1803.04042*, 2018.
- [21] Koh P W, Liang P. Understanding black-box predictions via influence functions[J]. *arXiv preprint arXiv:1703.04730*, 2017.
- [22] Bach S, Binder A, Montavon G, et al. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation[J]. *PloS one*, 2015, 10(7): e0130140.
- [23] Kim B, Khanna R, Koyejo O O. Examples are not enough, learn to criticize! criticism for interpretability[C]// *Advances in neural information processing systems*. 2016: 2280-2288.
- [24] Mehta P, Schwab D J. An exact mapping between the variational renormalization group and deep learning[J]. *arXiv preprint arXiv:1410.3831*, 2014.
- [25] Patel A B, Nguyen T, Baraniuk R G. A probabilistic theory of deep learning[J]. *arXiv preprint arXiv:1504.00641*, 2015.
- [26] Tishby N, Zaslavsky N. Deep learning and the information bottleneck principle[C]// *2015 IEEE Information Theory Workshop (ITW)*. IEEE, 2015: 1-5.
- [27] Ver Steeg G, Galstyan A. The information sieve[C]// *International Conference on Machine Learning*. 2016: 164-172.
- [28] Lu Y, Zhong A, Li Q, et al. Beyond finite layer neural networks: Bridging deep architectures and numerical differential equations[C]// *International Conference on Machine Learning*. 2018: 3276-3285.
- [29] Paul A , Venkatasubramanian S . Why does Deep Learning work? - A perspective from Group Theory[J]. *Computer Science*, 2014.

- [30] Wang D, Zhang M, Li Z, et al. Modulation format recognition and OSNR estimation using CNN-based deep learning[J]. *IEEE Photonics Technology Letters*, 2017, 29(19): 1667-1670.
- [31] Zhang J, Li Y, Yin J. Modulation classification method for frequency modulation signals based on the time–frequency distribution and CNN[J]. *IET Radar, Sonar & Navigation*, 2017, 12(2): 244-249.
- [32] Peng S, Jiang H, Wang H, et al. Modulation classification using convolutional neural network based deep learning model[C]//2017 26th Wireless and Optical Communication Conference (WOCC). IEEE, 2017: 1-5.
- [33] Lee J H , Kim K Y , Shin Y . Feature Image-Based Automatic Modulation Classification Method Using CNN Algorithm[C]// 2019 International Conference on Artificial Intelligence in Information and Communication (ICAIC). 2019.
- [34] Yongshi W , Jie G , Hao L , et al. CNN-based modulation classification in the complicated communication channel[C]// 2017:512-516.
- [35] Karanov B, Chagnon M, Thouin F, et al. End-to-end deep learning of optical fiber communications[J]. *Journal of Lightwave Technology*, 2018, 36(20): 4843-4855.
- [36] Schmitz J, von Lengerke C, Airee N, et al. A deep learning wireless transceiver with fully learned modulation and synchronization[C]//2019 IEEE International Conference on Communications Workshops (ICC Workshops). IEEE, 2019: 1-6.
- [37] Arvinte M, Vishwanath S, Tewfik A H. Deep learning-based quantization of L-values for Gray-coded modulation[C]//2019 IEEE Globecom Workshops (GC Wkshps). IEEE, 2019: 1-6.
- [38] Yang P, Xiao Y, Xiao M, et al. Adaptive spatial modulation MIMO based on machine learning[J]. *IEEE Journal on Selected Areas in Communications*, 2019, 37(9): 2117-2131.
- [39] O'Shea T J, Hoydis J. An introduction to machine learning communications systems[J]. *arXiv preprint arXiv:1702.00832*, 2017.
- [40] O'shea T, Hoydis J. An introduction to deep learning for the physical layer[J]. *IEEE Transactions on Cognitive Communications and Networking*, 2017, 3(4): 563-575.
- [41] Dörner S, Cammerer S, Hoydis J, et al. Deep learning based communication over the air[J]. *IEEE Journal of Selected Topics in Signal Processing*, 2017, 12(1): 132-143.
- [42] Salamon J , Jacoby C , Bello J P . A Dataset and Taxonomy for Urban Sound Research[C]// acm International Conference on Multimedia. ACM, 2014.
- [43] Xu B, Wang N, Chen T, et al. Empirical evaluation of rectified activations in convolutional network[J]. *arXiv preprint arXiv:1505.00853*, 2015.
- [44] Rönberg J, Lunner T, Zekveld A, et al. The Ease of Language Understanding (ELU) model: theoretical, empirical, and clinical advances[J]. *Frontiers in systems neuroscience*, 2013, 7: 31.
- [45] Trottier L, Gigu P, Chaib-draa B. Parametric exponential linear unit for deep convolutional neural networks[C]//2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA). IEEE, 2017: 207-214.
- [46] Goodfellow I, Warde-Farley D, Mirza M, et al. Maxout networks[C]//International conference on machine learning. 2013: 1319-1327.
- [47] G. Parascandolo, H. Huttunen, T. Virtanen. Taming the waves: sine as activation function in deep neural networks. *ICLR*, 2017.
- [48] Bracewell R N, Bracewell R N. *The Fourier transform and its applications*[M]. New York: McGraw-Hill, 1986.
- [49] Francesca M , Hughes A , Gregg D . Spectral Convolution Networks[J]. 2016.
- [50] Springenberg J T, Dosovitskiy A, Brox T, et al. Striving for simplicity: The all convolutional net[J]. *arXiv preprint arXiv:1412.6806*, 2014.
- [51] Glorot X, Bordes A, Bengio Y. Deep sparse rectifier neural networks[C]//Proceedings of the fourteenth international conference on artificial intelligence and statistics. 2011: 315-323.
- [52] Ioffe S , Szegedy C . Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift[J]. 2015.
- [53] Zeiler M D, Krishnan D, Taylor G W, et al. Deconvolutional networks[C]//2010 IEEE Computer Society Conference on computer vision and pattern recognition. IEEE, 2010: 2528-2535.
- [54] Dosovitskiy A, Tobias Springenberg J, Brox T. Learning to generate chairs with convolutional neural networks[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015: 1538-1546.
- [55] Deng J , Dong W , Socher R , et al. ImageNet: A large-scale hierarchical image database[J]. *Proc of IEEE Computer Vision & Pattern Recognition*, 2009:248-255.
- [56] Bainbridge, W. A. & Oliva, A. A toolbox and sample object perception data for equalization of natural images. *Data Brief*. 5, 846–851 (2015).
- [57] Ruderman D L. The statistics of natural images[J]. *Network: computation in neural systems*, 1994, 5(4): 517-548.
- [58] Srivastava A , Lee A B , Simoncelli E P , et al. On Advances in Statistical Modeling of Natural Images[J]. *Journal of Mathematical Imaging & Vision*, 2003, 18(1):17-33
- [59] Hou X , Zhang L . Saliency Detection: A Spectral Residual Approach[C]// IEEE Conference on Computer Vision & Pattern Recognition. IEEE, 2007.
- [60] Mesaros A , Heittola T , Virtanen T . TUT database for acoustic scene classification and sound event detection[C]// Signal Processing Conference. IEEE, 2016.
- [61] Everingham M, Winn J. The pascal visual object classes challenge 2012 (voc2012) development kit[J]. *Pattern Analysis, Statistical Modelling and Computational Learning*, Tech. Rep, 2011, 8.