# Unified Deep Learning

**Wang Wei**
weiwangnnn@yeah.net

## Abstract

With the increase in depth and complexity, deep learning networks have made significant progress. However, the theoretical understanding of deep learning remains incomplete. Early works in this field successfully demonstrated the Universal Approximation Theorem, but these proofs were limited to linear-type networks. This paper aims to bridge the gap in theoretical understanding of deep learning by proposing a unified approach. Surprisingly, all these network architectures (Linear, Convolutional, and Transformer) can be represented within this unified framework as a matrix multiplying a vector. Furthermore, this paper proves that multi-layer neural networks, including Linear, Convolutional, and Transformer networks, also satisfy the original approximation theorem. The key differentiating factor among these multi-layer networks lies in the form of parameters they learn. Linear networks learn dense matrices, while Convolutional networks and Transformers learn sparse matrices, which enhances their effectiveness in certain tasks and significantly reduces the parameter requirements. By establishing this unified framework and proving the applicability of the approximation theorem to multi-layer networks, this paper takes an important step towards unifying the entire field of deep learning. It deepens our theoretical understanding and reveals the fundamental principles governing the remarkable capabilities of these networks. It also paves the way for exploring new research avenues and optimizing the learning process in various deep learning applications.

## 1 Introduction

This article proposes a unified research approach for the theoretical study and model development of deep learning, providing valuable theoretical references. It demonstrates the unity of deep learning methods through a series of processes and proves the effectiveness of the deep learning approximation theory.

Firstly, the article offers a new perspective on the relationship between convolution, Transformer, and matrix-vector multiplication. This implies that both convolution and Transformer can be represented using the same form of matrix-vector multiplication. This unity provides a fresh view to understand the connections and transformations between different models.

Furthermore, the article proves that these forms of deep networks comply with the "universal approximation theorem," which is a crucial theoretical foundation of deep learning. The "universal approximation theorem" states that under certain conditions, a deep network can approximate any function to arbitrary precision. This result is essential for understanding the representational and learning capabilities of deep learning models, providing a solid foundation for the application and theoretical research of deep learning.

In summary, this article's contributions can be highlighted in two main aspects:

- 1. It proposes a unified approach for the fundamental modules of deep learning models - Linear, convolution, and Transformer - as well as a unified research scheme for matrix multiplication. This unity offers new insights into the connections and transformations between deep learning methods.

Preprint. Under review.

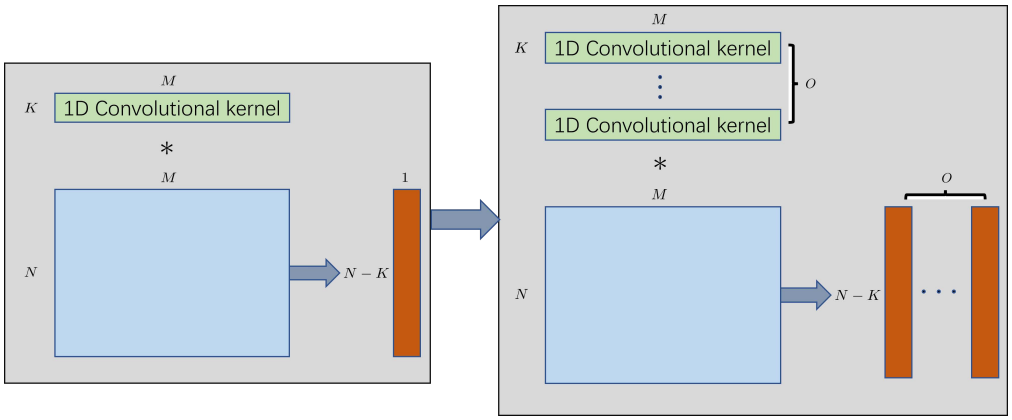


Figure 1: The diagram illustrates 1D convolution, with single-kernel and single-channel output on the left, and multiple-kernel and multiple-channel output on the right.

- 2. It proves that most deep networks constructed based on Linear, convolution, and Transformer conform to the "universal approximation theorem," reinforcing the position of deep learning as a powerful function approximation tool and laying a theoretical foundation for the development and application of deep learning.

## 2 Matrix-Vector

In this section, I will introduce my unified deep learning network concept. Firstly, a deep learning network consists of inputs, outputs, and various intermediate transformations. Regardless of the input's shape, the initial step involves treating each row as a column vector. These column vectors are then concatenated in the column direction. Subsequently, the intermediate transformation methods are represented in the form of sparse matrices. These sparse matrices are multiplied by the column vector to obtain the output of each layer, ultimately leading to the final output. The final output is then reshaped. This approach is designed to enable us to investigate the entire realm of deep learning in a unified manner. In the following, I will illustrate this approach through the fundamental modules of convolution and Transformer.

## 3 Convolution to Matrix Mutilplication

Convolution is one of the most commonly used basic units in deep learning. Therefore, this article will first demonstrate the relationship between convolution and matrix multiplication. The focus of this proof will be on 1D convolution, as the methods for proving 2D, 3D, or higher-dimensional convolutions are similar. Firstly, we present the simplest representation of 1D convolution in Figure 1.

Based on the representation of 1D convolution in Figure 1, we can write the matrix multiplication corresponding to the left-hand side single-channel convolution output in Figure 1, as shown in Figure **??**.

Furthermore, we can write the matrix multiplication corresponding to the right-hand side multi-channel convolution output in Figure 1, as shown in Figure 3.

Based on the above representation, we can express 1D convolution as a matrix multiplication with a vector. In formulaic form:

$$x^{i+1} = Wx^i \tag{1}$$

where $x^i \subset R^{M*N}$ and $x^{i+1} \subset R^{(N-k+1)*O}$. $W \subset R^{(N-k+1)*MN}$ is a sparse matrix generated according to the convolution kernel. Additionally, the proof approach for 2D and 3D networks is

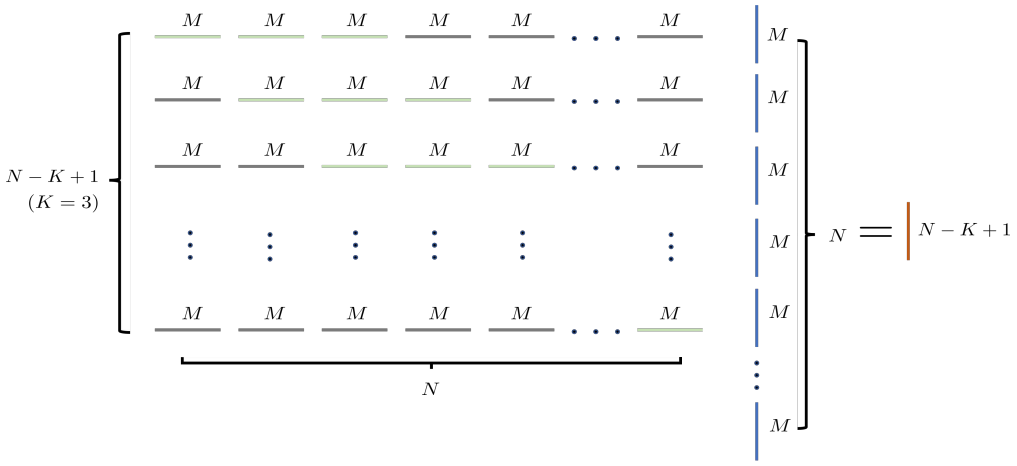


Figure 2: The matrix-vector representation of 1D convolution involves taking each row of the $N * M$ input matrix as a column vector, then concatenating $N * M$ column vectors in the column direction to form a new large column vector. Simultaneously, each row of the convolutional kernel is treated as a green row vector, and these three vectors are concatenated in the row direction. Next, arrange these vectors and matrices according to the layout shown in the diagram, filling the positions with grey vectors as zeros. This approach can get the output result of the convolution.

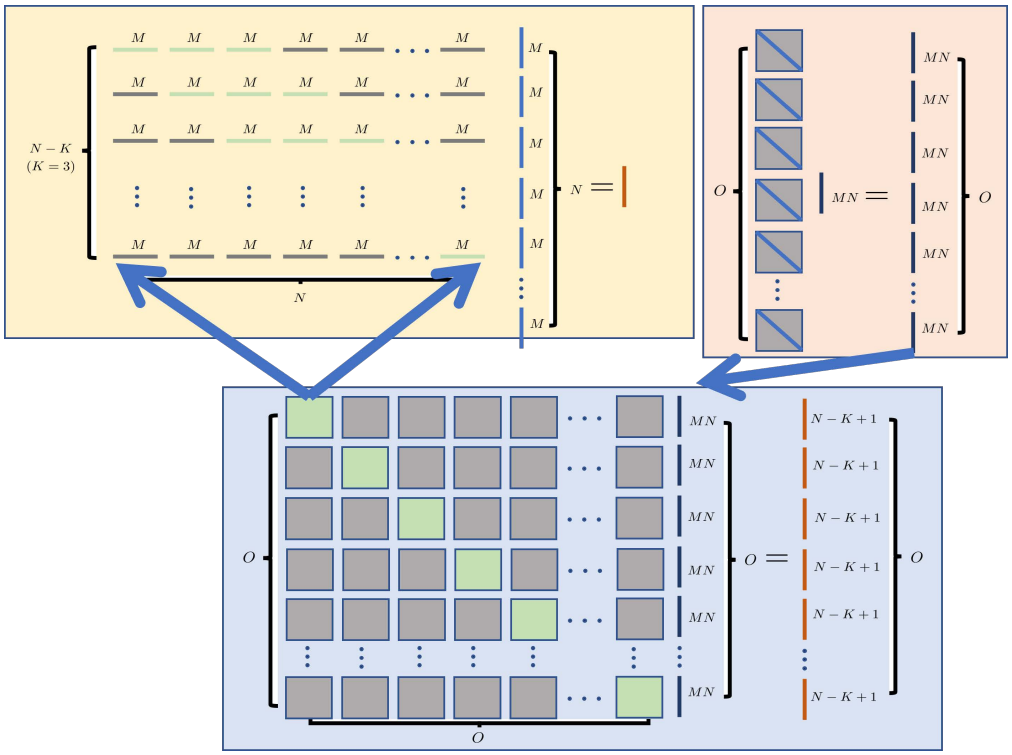


Figure 3: This is the matrix representation of convolution with multiple channel outputs. Firstly, the concatenated $M * N$ column vectors are copied $O$ times in the column direction using the identity matrix. Then, the convolution is written as a matrix multiplied by a column vector using the method described above. The representation in the diagram is the same as in Figure 2.

similar, and there is no need for further elaboration. Thus, we can represent convolution as matrix multiplication in this manner.

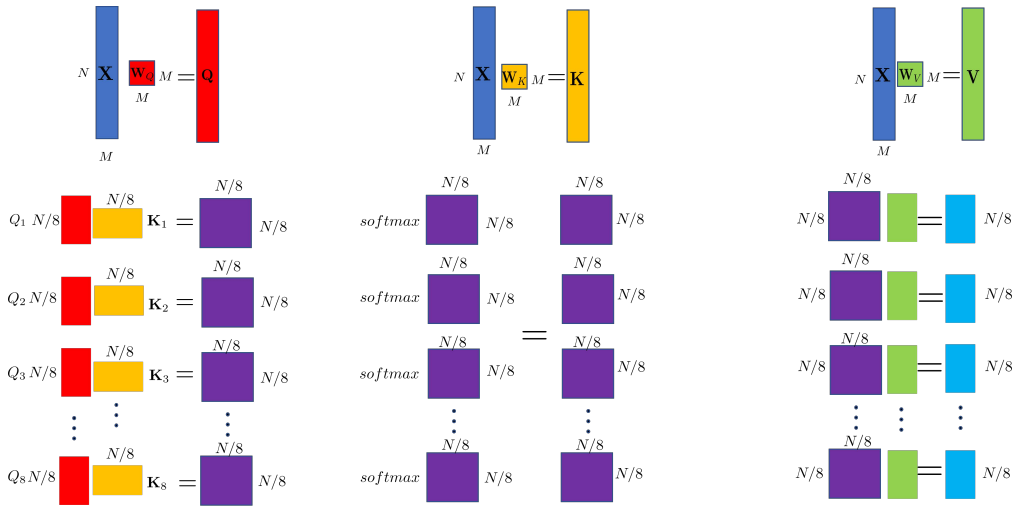


Figure 4: The image depicts the basic module of the Transformer, which is multi-head attention.

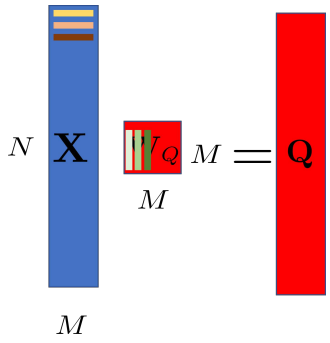


Figure 5: The image illustrates the concept of $x * W_Q$ in the multi-head attention of the Transformer. We represent different parts of the vectors in $x$ and $W_Q$ using different colors.

## 4 Transformer to Matrix Mutilplication

We have already expressed convolution in the form of a matrix multiplying a vector. In this section, we will continue with the same idea and represent the Transformer in a similar form. First, we present the general form of the Transformer, as shown in Figure 4

To explicitly show the matrix representation of the Transformer, we show the $XW_Q$ in Figure 5.

In Figure 6, we will rewrite $XW_Q$ in Figure 5 in the form of a matrix multiplying a vector, while carefully preserving the corresponding vector colors' positions in both images.

$XW_K$ and $XW_V$ are similar to the above. We define $W_{QK} = softmax(QK)$, which can be understood as a sparse weight matrix because it learned from $W_Q$ and $W_K$ and can be deducted according to the above idea, then represent $W_{QK}V$ in Figure 7.

Now, we know the Transformer can also be represented as a matrix multiply a vector which can be written as:

$$\begin{aligned} x^{i+1} &= W_{QK} W_V x^i \\ &= W_{QKV} x^i \end{aligned} \tag{2}$$

where $x^i \subset R^{N*M}$ and $x^{i+1} \subset R^{N*M}$. $W_Q K \subset R^{N*M}$ and $W_V \subset R^{N*M}$ are sparse matrixs generated according to Transformer.

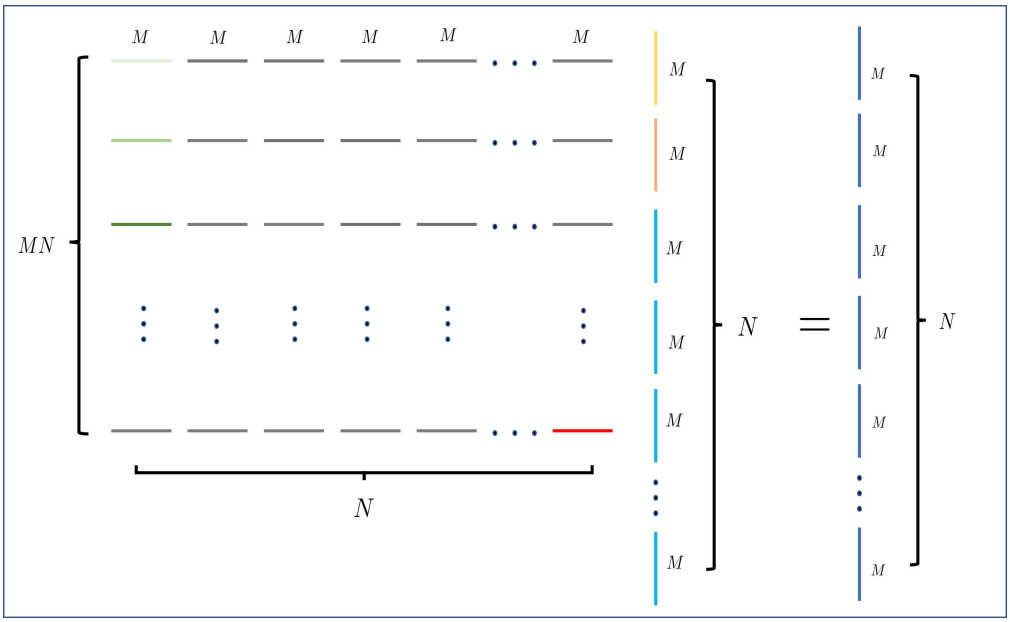


Figure 6: The representation of $x * W_Q$ in the multi-head attention by a matrix multiplying a vector .

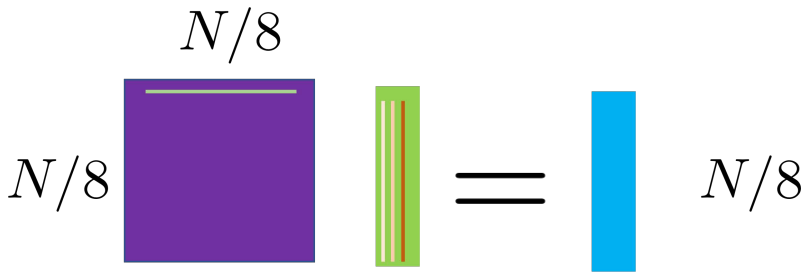


Figure 7: The representation of $W_{QK}V$. We use different color to represent different vectors in $W_{QK}$ and $V$

## 5 Universal approximation theory and a Matrix Times a Vector

Based on the above derivation, we have shown that both convolution and Transformer can be expressed as a matrix multiplied by a vector. In this representation, the matrix corresponds to a sparse parameter matrix, and the vector corresponds to the input matrix unfolded along the column direction. In this section, we will prove that any network composed of convolution and Transformer and structured in the form of $T_1(X) + \sigma(T_2(X))$ concatenation is also subject to the Universal Approximation Theorem, where $T_1$ and $T_2$ are convolution and Transformer or other Transformation which can be translated in the way of our proposed that a matrix multiplying a vector.

[1, 2] proposed the Universal Approximation Theorem. I used the form in [2]. $\sigma$ is any continuous sigmoidal function. Then finite sums of the form

$$G(x) = \sum_{j=1}^{N} \alpha_j \sigma\left(y_j^{\mathrm{T}} x + \theta_j\right)$$

are dense in $C\left(I_n\right)$. In other words, given any $f \in C\left(I_n\right)$ and $\varepsilon > 0$, there is a sum, $G(x)$, of the above form, for which

$$|G(x) - f(x)| < \varepsilon \quad \text{for all} \quad x \in I_n.$$

That is the conclusion in [2]. I will prove convolution and Transformer also fulfils that. We know convolution and Transformer and any transformation which can be written like $x^{i+1} = W^i x^i$

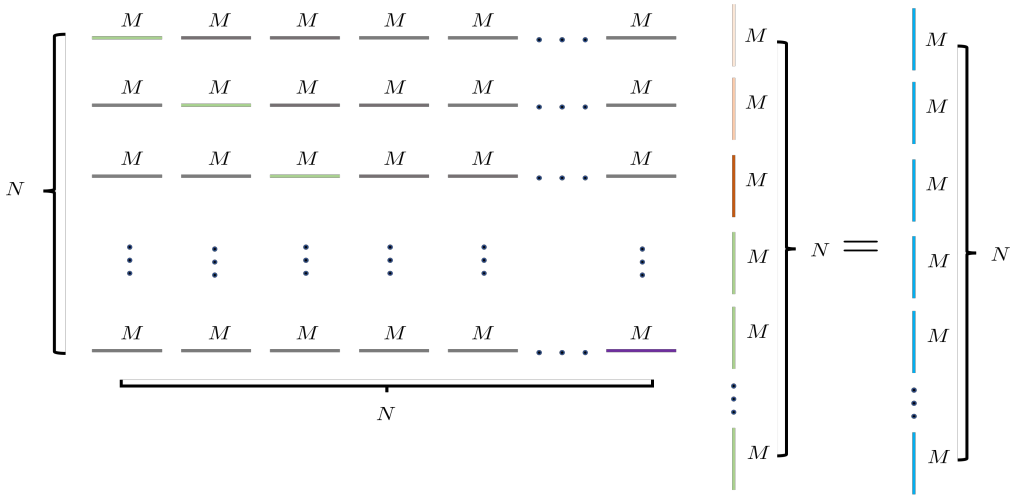


Figure 8: The matrix multiplication representation of $W_{QK}V$ in the form of a matrix multiplying a vector. Please pay attention to the correspondence between the vector colors in this figure and the vector colors in Figure 7.

according to our idea, where $x^i$ and $x^{i+1}$ represent the input and output of $i$th layer and $W^i$ are the sparse matrix deduced from transformation. We take a three layers Resnet as an example as following:

$$
\begin{aligned}
1\ layer :& W_1X + \sigma(W_1'X) \\
2\ layers :& W_2[W_1X + \sigma(W_1'X)] + \sigma\Big(W_2'[W_1X + \sigma(W_1'X)]\Big) \\
=& W_2W_1X + W_2\sigma(W_1'X) + \sigma\Big(W_2'W_1X + W_2'\sigma(W_1'X)\Big) \\
3\ layers :& W_3\Big[W_2[W_1X + \sigma(W_1'X)] + \sigma\Big(W_2'[W_1X + \sigma(W_1'X)]\Big)\Big] + \\
& \sigma\Big\{W_3\Big[W_2[W_1X + \sigma(W_1'X)] + \sigma\Big(W_2'[W_1X + \sigma(W_1'X)]\Big)\Big]\Big\} \\
=& W_3W_2W_1X + W_3W_2\sigma(W_1'X) + W_3\sigma\Big(W_2'W_1X + W_2'\sigma(W_1'X)]\Big) + \\
& \sigma\Big\{W_3'W_2W_1X + W_3'W_2\sigma(W_1'X) + W_3'\sigma\Big(W_2'W_1X + W_2'\sigma(W_1'X)\Big)\Big\}
\end{aligned}
\tag{3}
$$

It is obvious that it has the same format as Universal Approximation Theorem.

## 6 Conclusion

We have proposed a matrix-vector scheme that unifies the entire field of deep learning. Any transformation that can be expressed in this form shares similar mathematical characteristics with a single-hidden-layer linear neural network. The key distinction lies in the fact that while the weight parameters of a linear network are dense, the weight parameters of other transformations are sparse. For certain specific problems, such as image-related issues, these multi-layer sparse parameter matrices often aid in the network's convergence. Similar to convolutions and Transformers, they can also help reduce computational complexity. Nevertheless, from a theoretical standpoint, we can undertake a unified study of these diverse transformations.

## References

[1] Kurt Hornik. Approximation capabilities of multilayer feedforward networks. *Neural Networks*, 4(2):251–257, 1991.

[2] George Cybenko. Approximation by superpositions of a sigmoidal function. math cont sig syst (mcss) 2:303-314. *Mathematics of Control, Signals, and Systems*, 2:303–314, 12 1989.