Proceedings of the 5th International Conference on Computing and Data Science
DOI: 10.54254/2755-2721/20/20231079

Transformer model: Explainability and prospectiveness

Gengpan Nuobu

Beijing Normal University, Beijing , China

776636574@qq.com

Abstract. The purpose of Artificial Intelligence(AI) is to simulate learning process of human brain by strong computing power and appropriate algorithm, so that the machine can develop judging ability at work as human. Current AI mainly relies on Deep Learning model which is based on artificial neural network, like Convolutional Neural Network(CNN) in computer visualization, but that also takes with some defects. This paper introduces defects of CNN and discusses Transformer model in solving unexplainability of traditional CNN algorithm. To discuss why the Transformer model and attention mechanism are considered as the way to "AI intelligibility".

Keywords: deep learning, convolutional neural network(CNN), artificial neural network, transformer model attention mechanism.

1. Introduction

Computing the trace of cannonball was the purpose why the people invented computer, at the beginning of computer history, computer was used to do the single calculation work. Computer scientists never stop to develop other functions of the computer, trying to let the computer do the work more than a single calculation. After 21st century, people start to expect the computer can simulate the thinking process of human even better than human and work on it, instead of only let computer do the well-defined repetitive tasks, which means people want the computer technology evolve from programming intelligence to learning intelligence. Like figure 1. AI AlphaGo won the game of Weiqi in March 2016 and face recognition technology in daily life, are two examples of a machine can simulate human thinking. Use the algorithm to learn, analyze, predict, modify, control and finish the work that human can finish or can not.

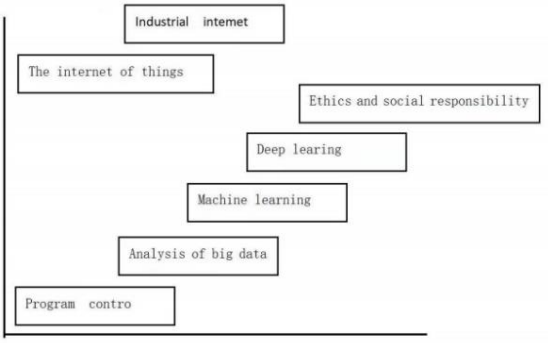


Figure 1. The evolution of AI.

© 2023 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/20/20231079

Machine learning technology is to study how to simulate the process of human thinking by computer, machine learning focus on the mechanism of a machine to learn new knowledge new skills and distinguish with those already existed. Machine learning modifies program algorithm by learning experience, using data and former experience to optimize the performance of computer applications standard, such as go a better step in weiqi or improve accuracy of face recognition, improving ability of the algorithm in working and learning.

2. Literature review

Machine learning has nearly 70 years of history, it has many good models and algorithm, neural network learning is one much more successful than others, which is also called Deep Learning. The concept of Deep Learning was first mentioned by G.E.Hinton in the University of Toronto in 2006, Deep Learning is a kind of deep network struct ma- chine learning process which is based on sample data and training [1,2]. Deep Learning process can be explained as simulating the neural network of the human brain made up of neurons by algorithm model, but in applications it is much more complicated than that explanation. Mostly, Deep Learning uses multilayer neural network which has excellent feature learning ability. Multilayer neural network is able to learn more essential in sample data. Multilayer neural network need multilayer understand input treatment when it dealing with complex problems.

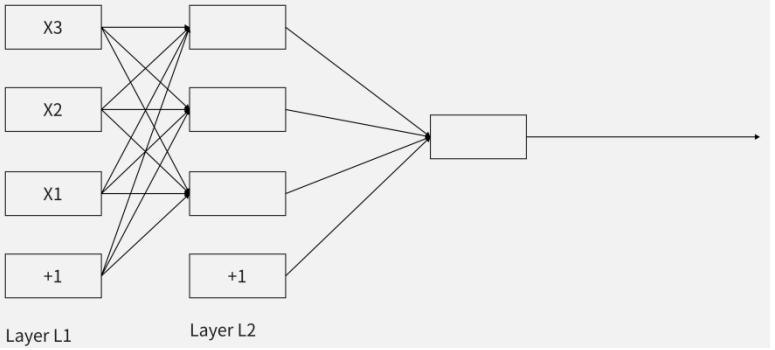


Figure 2. Process of an artificial neural network.

The process of artificial neural network is like figure 2 The number of neurons per layer. The number of layers and cumulative learning and training time determines the advantages and disadvantages of deep learning. For example, AI AlphaGo has 13 layers and 192 neurons in per layers, and that support it to learn the Weiqi game much better than most of the human player.

Though AI can play the Weiqi game very well, it does not mean AI have not defects. There were several ups and downs in the AI field. Every time is caused by a solution of the main problem or some important defects of AI. If the defects can be overcome, then it is much easier to take a period of rapid development to AI [3].

This paper wants to introduce current defects of Deep Learning model, so the work I have done was read many lecture review of deep learning model, collect the defects different researcher think and organize their ideas. There are three main defects in current Deep Learning model.

2.1. Unavailable useful data.

For Deep Learning model, plenty of data with label is of great importance, but in fact, that kind of data is hard to get. Because data collection usually consumes a lot of resources. In addition, privacy and safety of data is paid more and more attention recently, which make data collection harder [4]. Thirdly, data is needed to be marked by human. This work will cost large labor cost and time. Thus, data with label which Deep Learning truly relies on being difficult to obtain in fact.

DOI: 10.54254/2755-2721/20/20231079

2.2. Computing power

The foundation of Deep Learning is a neural network algorithm, if number of neurons per layer and number of layers increase, the computational time of the computer will rapidly increase which means the cost will increase too [3-6]. Let AlphaGo as example, the computing power of it is 11.38GFLOPS, computing power cost of a computer is 35,000,000 dollars. In modern society, it is not just colleges and universities, profitable companies are also trying to develop Deep Learning model, high computing power cost is easily restricted rapid development of Deep Learning.

2.3. Unexplainability

Deep Learning model mostly focusses on the relationship between learning data and label, but rarely focusses on physical meaning [4]. In other words, even Deep Learning model itself can not explain what it has learned and few of theories can verify or explain what can a Deep Learning model can learn finally, or how much it can learn? Like the founder of Bayesian networks Pearl said “Almost all the deep learning in essence is curve fitting”. Process of Deep Learning is hardly can be explained. Usually is considered as a black box model by relevant workers [4].

2.4. Other problems

2.4.1. Ethical issues. If AI will exist with human, working in human society. Then AI should take the relevant responsibility [6]. Imagine a scene like this, AI control a car that does not work right, have to kill passengers or passerby to stop the car, then what will the AI do, who will take the responsibility? How to take the responsibility? Take the AI in to prison?

2.4.2. Trust issues. Because the Deep Learning model is still unexplainable, people can not explain and predict what and how much will Deep Learning learn, thus human can not trust AI. Person can trust a kitchen knife he or she using because everything is explainable and can be predicted as long as kitchen knife is controlled, but Deep Learning model is not as explainable as human need, so trust issue exists.

3. Methodology of CNN

3.1. Math of CNN

It is obvious that if we want to know the advantage of Transformer model, We better know the methodology of CNN so that we can also know why. As mentioning before, CNN is mostly used in computer visualization. If a person sees picture , he can easily know the color, curve, shape, which can help him quickly understand what the picture wants to express. But that is truly difficult on a computer which can only use a number matrix to describe a picture.

RGB color mode is widely used in computer visualization, it is using red(R), green(G) and blue(B) to display the color, each intensity is between 0 and 255. If we have a figure with size of 480×480, the input of a computer is an array with size of 480×480×3. The elements of an array are the color intensity of position.

Like figure 3, assume that there is a figure input with 8×8×3 CNN algorithm has a neuron every time calculate a small part of input number matrix(in figure3, the neuron can calculate 3×3) like flashlight light up a small place in one time. Figure3 is the process of first layer of CNN, which is also named as convolution layer, by this process, an input array size of 8×8×3 will be calculated an there will be a 6×6×3 array output, and this array with size 6×6×3 is also the input of second layer, like figure 2.

DOI: 10.54254/2755-2721/20/20231079

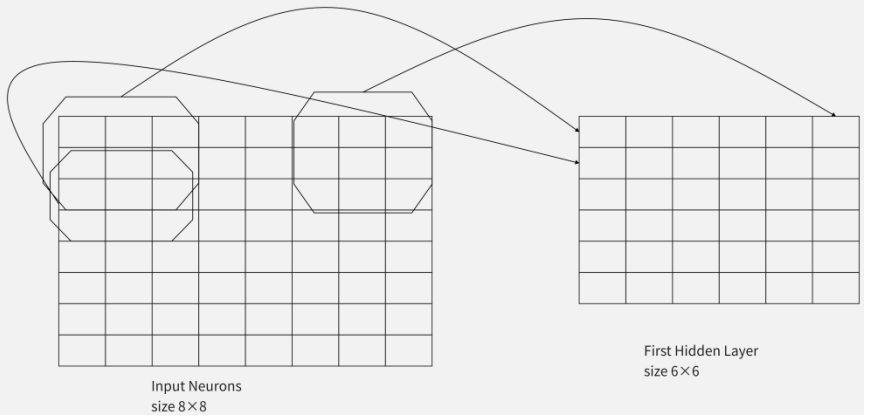
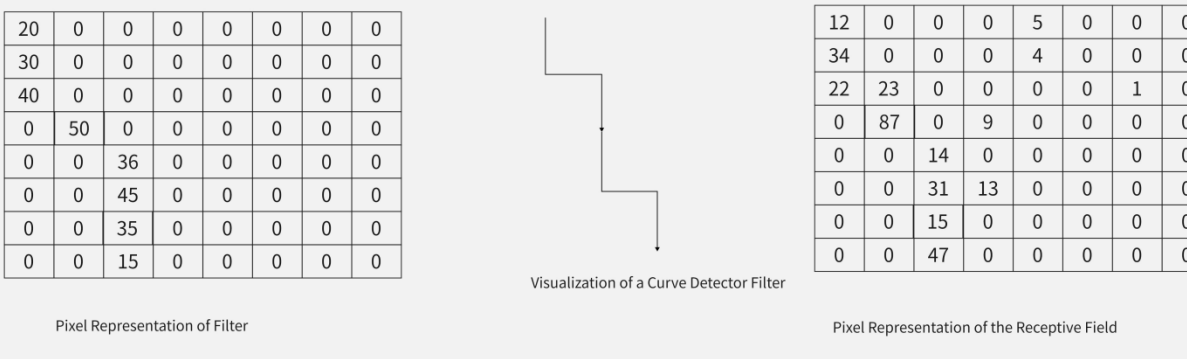


Figure 3. Convolution of CNN input.



20	0	0	0	0	0	0	0
30	0	0	0	0	0	0	0
40	0	0	0	0	0	0	0
0	50	0	0	0	0	0	0
0	0	36	0	0	0	0	0
0	0	45	0	0	0	0	0
0	0	35	0	0	0	0	0
0	0	15	0	0	0	0	0

12	0	0	0	5	0	0	0
34	0	0	0	4	0	0	0
22	23	0	0	0	0	1	0
0	87	0	9	0	0	0	0
0	0	14	0	0	0	0	0
0	0	31	13	0	0	0	0
0	0	15	0	0	0	0	0
0	0	47	0	0	0	0	0

Figure 4. First layer process.

In figure 4, there is a curve supposed to be recognized. Pixel Representation of Filter is an array of curve which is also the neuron to recognize this curve. Pixel Representaion of Receptive Field is input of the first layer. What will happen if two arrays dot product? If a figure is similar to this neuron, the dot product will be extremely large. In figure 3.2 the dot product is 9619(02×12+30×34+40×22+50×87+36×14+45×31+35×15+15×47=9619). On the contrary, if the figure is not included curve the neuron wants, the dot product will be very small. Like neure receive information and finally output a positive or negative information in the nervous system, neuron of the first layer can differentiate the large result and small result, to know whether the input figure include curve. One neuron finds a line, with the neuron number increase. CNN can identify more lines and those will be second layer's input.

As for deeper layers, the mathematical process is similar to the first layer, but it is hard to describe with visualization that human can understand easily. The last layer is of great importance, which is also called Fully Connected Layer. Fully Connected Layer output a vector of possibility of each expected result. Use roman numerals as an example, it has twelve numbers, that makes the Fully Connected Layer will output a normalized 1×12 array like [a1, a2, a3, a4, a5, a6, a7, a8, a9, a10, a11, a12], the possibility of input figure express roman numeral 1 is a1, and so on.

3.2. Training of CNN

As we mentioned before, a neural network needs to be trained on a large number of data on label. With labeled data, CNN will repeatedly modify the parameters and finally high accuracy can be achieved.

In the training, the difference of output and expected target might be huge, in that case the parameter needs to be modified. In equation 1, L is named cost function

$$L = \sum \frac{1}{2}(\text{target} - \text{output})^2 \quad (1)$$

DOI: 10.54254/2755-2721/20/20231079

$$w = w_i - \eta \frac{dL}{dW} \quad (2)$$

In equation 2, w is the weight, w_i is initial weight, and η is learning rate. Parameter on CNN will be modified like equation 2.

4. Transformer model

Unexplainability is a very important theoretical problem in traditional CNN. There were some work done tried to solve the unexplainability, but the explanation effect is not as good as expected lack unified evaluatation index explainability application fields were limited [7].

There were many solutions tried to solve the unexplain-ability, like neural network algorithm base on explainable module trying to extract some logical rules or pattern which is easy to understand from trained networks upgrade the explanation by input human knowledge or build model with human centered [8-10]. Recent main studys in explainability of CNN are mostly focus on the explanation of the behavior and judgement of model, but it is hard to let a CNN model become explainable without damaging the performance of network [11-13].

Based on the above reasons, some researchers now are trying to make the Deep Learning model explainable with attention mechanism. Attention is one of the largest research fields in cognitive science. Attention map and dynamic attention mechanism are two of important features in attention of the human brain. Attention map has two kinds, Saliency map and Priority map [5].

Based on the perspective of bionics, recent Deep Learning is constantly adopting and studying attention relevant research result in cognitive science to improve Deep Learning especially trying to figure out the explainability problem, and attention mechanism is also considered as a significant way to "explainable AI", Transformer model is one of the results trying to combining attention mechanism and Deep Learning.

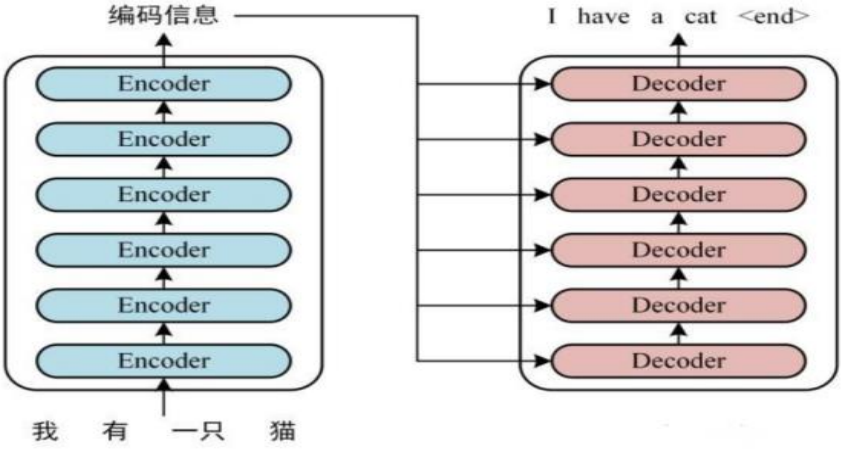


Figure 5. Translation diagram of transformer model.

Transformer model is firstly used on treatment of natural languages. Transformer model used an encoder-decoder structure which can transform input features to Query key value, totally three matrices, calculate every attention weight of input words with Query matrix and key matrix. Using self attention weight to weight sum the value can get a special relationship feature output between every word, like the figure 5, is the process of languages translation. Not only natural language processing, Transformer model can also get good results in many other fields such as computer visualization. Like figure 6, the main module of Transformer model is attention module, which is a score of all positions to calculation similarity in input array.

DOI: 10.54254/2755-2721/20/20231079

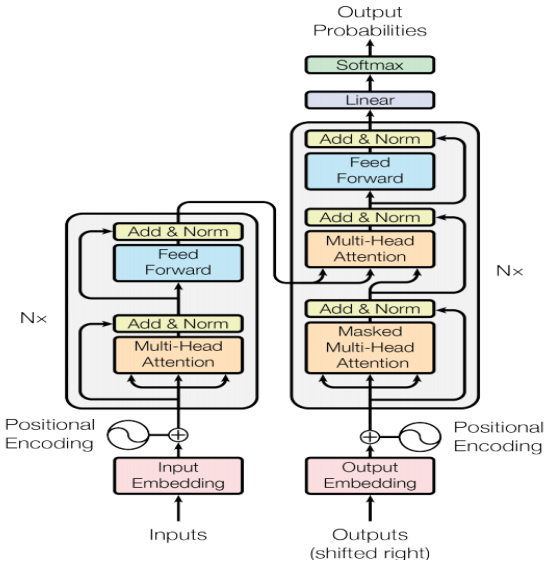


Figure 6. The Transformer - model architecture.

October 2020, Lambda Networks, a kind of Transformer model is proposed, which use a Lambda layer to provide captured input and a group of structured frameworks between context elements to express networks model, is considered as attention mechanism replacement. In other words, Lambda layer defines the similarity of nuclear between elements and context elements, summary of the context information as a linear function.

5. Methodology of Transformer model

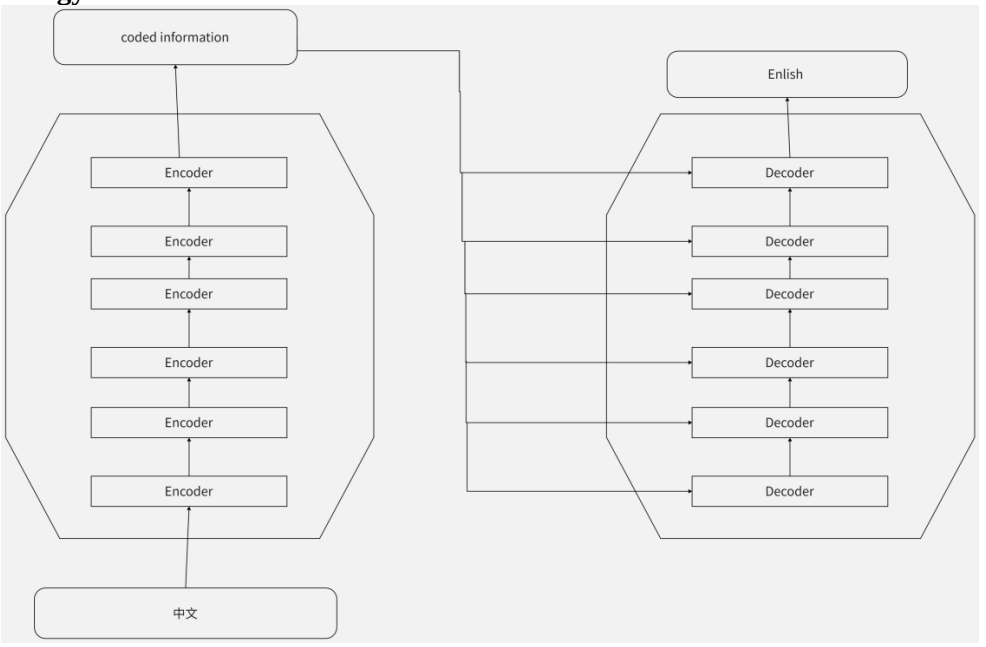


Figure 7. Structure of Transformer model.

Figure 7 is the structure of Transformer model, its work includes two main processes, encode process and decode process, which also let its structure can be built in two parts, encoder part and decoder part. We will introduce the input of Transformer model at the beginning of introducing the mechanism of Transformer model.

DOI: 10.54254/2755-2721/20/20231079

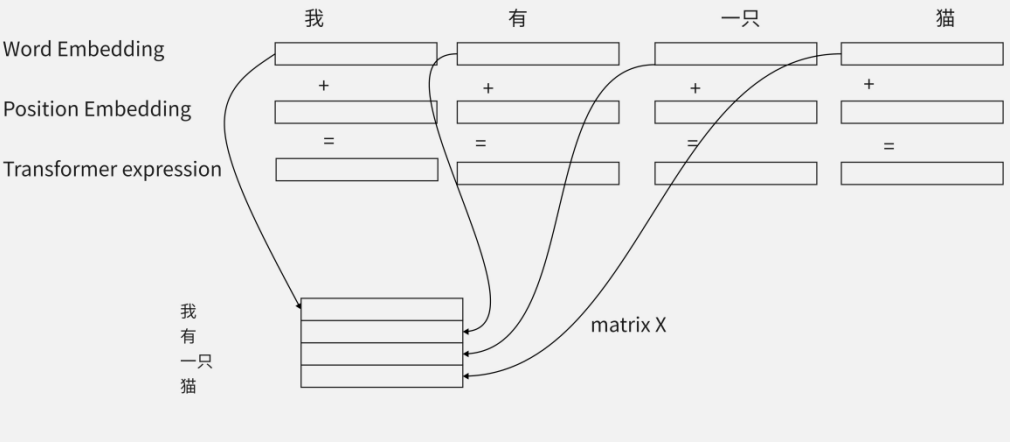


Figure 8. The input of Transformer model.

5.1. Input of Transformer model

The input of Transformer model is like figure8. Every word in a sentence is a vector which is gotten by two vectors added together, one vector is Word Embedding(WE) and another one is Position Embedding(PE). The method to obtain WE is diverse training in Transformer model or using other algorithm. PE includes the position information of every word in a sentence, mostly, we do not care about its absolute position but relative position. For that reason, Trans- former model uses the equation to calculate the PE like in equation 3, which is more convenient to know the relative position instead of absolute position.

$$\begin{cases} \mathrm{PE}_{\mathrm{POS},2i} = \sin\left(\frac{\mathrm{pos}}{10000^{\frac{2i}{d}}}\right) \\ \mathrm{PE}_{\mathrm{POS},2i+1} = \cos\left(\frac{\mathrm{pos}}{10000^{\frac{2i}{d}}}\right) \end{cases} \tag{3}$$

In equation 3, pos is the relative position of a word and d is the dimension of PE. 2i+1 means an odd number of dimensions and 2i mean an even number of dimensions ($0<2i<d, 0<2i+1<d$). There are two advantages for us to use an equation like equation 3 to calculate the PE instead of training by some algorithm.

•We can use the equation to calculate the PE even we input a very long sentence, the only thing different is the number.

•Like the equation 4 , the relative position of two words can be easily calculated if PE is obtained by using the equation3

$$\begin{cases} \sin(\alpha+\beta) = \sin\alpha\cos\beta + \cos\beta\sin\alpha \\ \cos(\alpha+\beta) = \cos\alpha\cos\beta - \sin\alpha\sin\beta \end{cases} \tag{4}$$

By this way, we will get an input matrix $X_{n\times d}$, n is the number of words in sentence and d is the dimension of input vector. In paper Attention is All you Need [14], d=512.

5.2. Encoder block

After we have input matrix X, X will enter the encode part, which is made up of six encoder blocks. Structure of one encoder block is like figure 9.

And the structure of Multi-Head Attention is like figure 10, in another word, Multi-Head Attention block is a linear transformation of many Self-Attention Blocks, Like figure 11, it can be described as linear transformation of many Self-Attention layers.

DOI: 10.54254/2755-2721/20/20231079

Every Self-Attention Block has three linear transformation matrices WQ, WK and WV, we assume the input of Self-Attention Block is $X_{n\times d}$, then firstly what we have is three key matrices Q , K and V. In paper Attention is All you Need [14], size of matrix WQ, WK and WV is 512×64, size of Q, K and V is same as X, is n×512. In equation 5, d_l=64 , d_k=512.

$$\begin{cases} Q_{n\times d_k} = X_{n\times d_l} \cdot WQ_{d_l\times d_k} \\ K_{n\times d_k} = X_{n\times d_l} \cdot WK_{d_l\times d_k} \\ V_{n\times d_k} = X_{n\times d_l} \cdot WV_{d_l\times d_k} \end{cases} \tag{5}$$

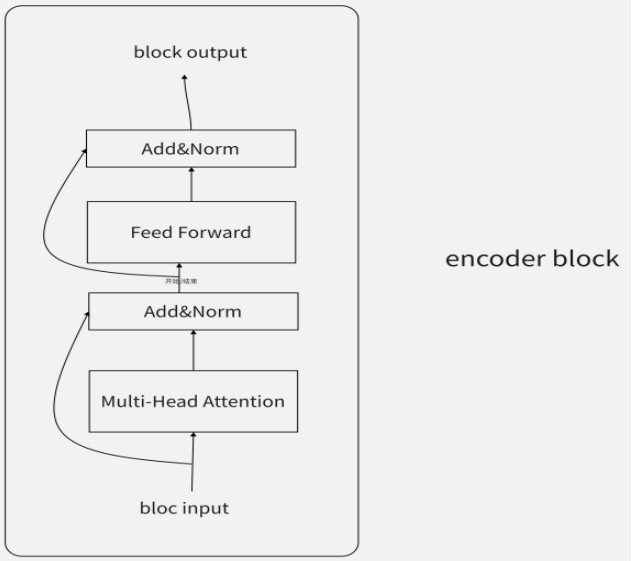


Figure 9. Structure of encoder block.

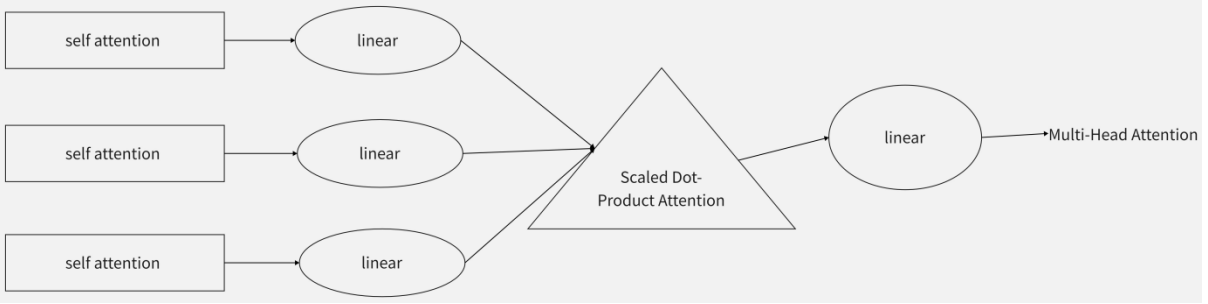


Figure 10. Structure of Multi-Head Attention.

And the structure of Multi-Head Attention is like figure 10, in another word, Multi-Head Attention block is a linear transformation of many Self-Attention Blocks, Like the figure 11, it can be described as a linear transformation of many Self-Attention layers.

We calculate the output of Self-Attention Block's output Z by equation 10. The softmax is an operator that let the sum of every line of matrix to be positive one, which can decrease the difference of every element, is helpful to fast convergence for computational processes and normalization. The concept of Q, K and V is an information retrieval system, if you input computer, Q is a computer, it will find all the things that have high similarity with what you input. In other word, search engine search highly similar things K with Q, and search engine will match the V by similarity between Q and V. In transformer model, matrix multiplication is the method to find the similarity of Q and K, and weighted sum is used to calculate similarity of Q and V.

$$Z_{n\times d_k} = \text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d_l}})V \tag{6}$$

DOI: 10.54254/2755-2721/20/20231079

Like figure 11 and figure 10, the output matrices Z_i will be linear transformed and be output of Multi-Head Attention.

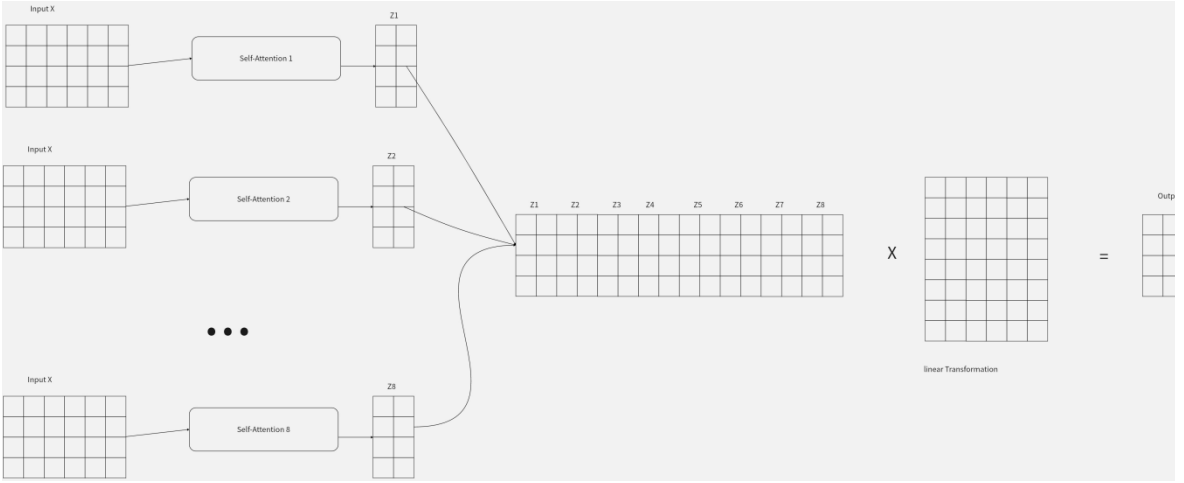


Figure 11. Process of Multi-Head Attention.

The Add&Norm part has two subparts, Add part is a kind of residual connection that used to let neural network concentrate on present difference. The Norm part is layer Normalization that can be used to accomplish the rapid convergence.

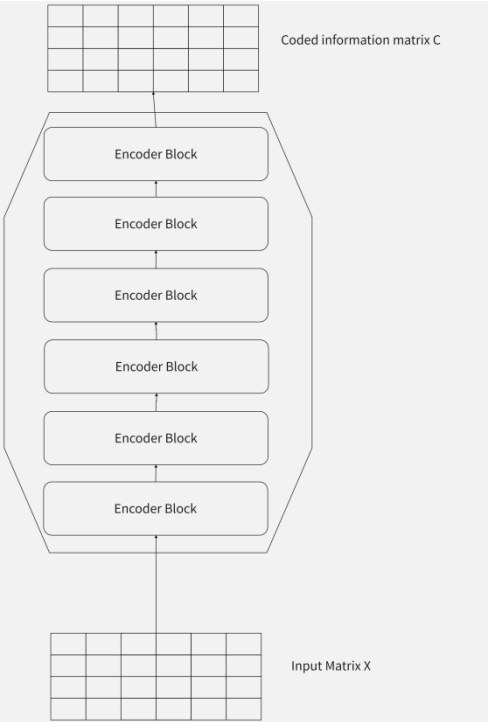


Figure 12. Process of Calculate Coded Information Matrix C.

Thus, the input matrix X will go through six Encode Blocks and finally produce Coded Information Matrix C as output of encoding part of Transformer model.

5.3. Decoder Block and output of transformer model

The structure of Decoder Block ls like figure 13. In one Decoder Block, Masked Multi-Head Attention is different with Multi-Head Attention we mentioned before is that is used to Masked operation.

DOI: 10.54254/2755-2721/20/20231079

Because the sentences input is word by word and words have it's sequence. Translation can not ignore the sequence, so using masked operation can cover the words after the one now translating. Except that, Masked Multi-Head Attention is same as what we mentioned in Encoder Block.

The Multi-Head Attention in Decoder Block, matrix K and V is not from the output of last Decoder Block, they are from Coded Information matrix C of Encoder Blocks, Q is from output matrix Z of last Decoder Block.

Output of six Decoder Blocks is a final output matrix Z. Z will be operated by softmax to predict every word like figure 14.

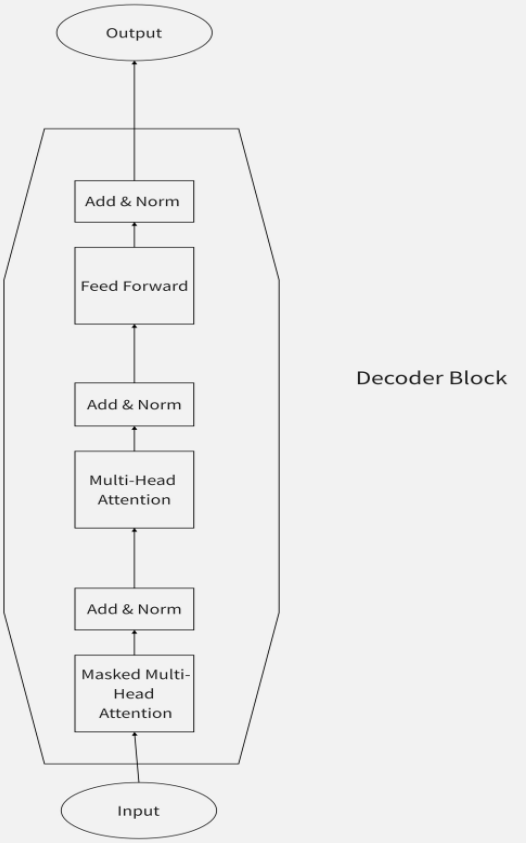


Figure 13. Structure of Decoder Block.

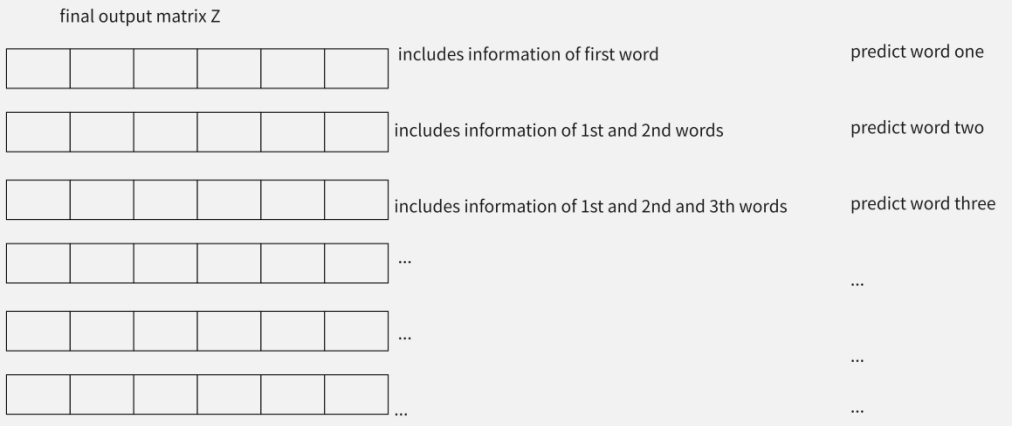


Figure 14. Process of Word Prediction.

DOI: 10.54254/2755-2721/20/20231079

6. Result

6.1. Advantage of transformer model

1. Transformer model abandons the CNN algorithm and already has some pretty good achievement, it's design is innovative and is worthy to learn. It is possible to solve the unexplainability problem because it abandons CNN.

2. Transformer model set distance of every two words as 1, which can improve performance.

3. Transformer model can be widely used, not only natural language processing, but also can work well in other fields like computer visualization.

4. Parallelism of the algorithm is much better than traditional CNN.

6.2. Disadvantage of transformer model

5. Simple Transformer model without CNN or RNN has no ability of capturing local features.

6. Transformer model loss position information, which is very significant in natural language processing. Thought Transformer model uses PE, it does not improve the defect.

7. Conclusion

In this paper, we introduced the purpose to develop AI is trying to simulate the process of human thinking by Deep Learning model so that computer scientists are able to build a machine to start learning and do the work even better than human. But Deep Learning has some defects restrict rapid development of Deep Learning. Unexplainability is one of the most important one. Recently, researchers start combining cognitive science and AI to solve the unexplainability of deep learning, combining the attention mechanism and Deep Learning model to make the neural network become more explainable.

The future, at least I wish , is to solve the unexplainability of Deep Learning, let human trust AI and become willing to work with AI, in that case, AI will provide a great development of social productive forces and technology will once again benefit mankind to a great extent.

References

[1] HINTON G E OSINDE O S TEH Y W A fast learning algorithm for deep belief nets J Neural Computation 2006 18(7):1527-1554.

[2] BENGIO Y Learning deep architectures for AI J Foundations and Trends in Machine Learning 2009 2(1):1-127.

[3] Yue Hu, Dongyang Luo, Kui Hua, Haiming Lu & Xue- gong Zhang. (2019). Review and discuss about deep learning. Journal of Intelligent Systems (01), 1-19.

[4] Jiangpeng Wang & Cheng Wang. (2021). Process and future trend of deep learning technology. China's Security(07),81-84.

[5] Yakun Wang. (2021). Development review of deep learning technology in 2020. Unmanned system technology(02),1-7. doi:10.19942/j.issn.2096- 5915.2021.2.011.

[6] Weimin He, Zheng He & Na Zhou. (2021). Hot and cold of deep learning. Microcontroller and embedded system application(11),2-4.

[7] Yingying Hua,Daichi Zhang & Shiming Ge. (2020). Study progress of Deep Learning model's explainability. Journal of Information Security (03),1-12. doi:10.19363/J.cnki.cn10-1380/tn.2020.05.01.

[8] FeiWu, BinBing Liao & Yahong Han. (2019). Explainability of Deep Learning. Aviation weapon (01), 39-46.

[9] Lingmin Li, Mengran Hou, Kun Chen & Junmin Lu. Review of Deep learning's explainability study. Computer application.

[10] Keyang Cheng, Ning Wang,Wenxi Shi & Yongzhao Zhan. (2020). Study progress of Deep learning's explain-ability. Computer Study and Development(06),1208-1217.

DOI: 10.54254/2755-2721/20/20231079

[11] Siming Yang, Zheng Shan, Yu Ding & Gangwei Li. (2021). Review of deep reinforcement learning study. Computer Engineering(12),19-29. doi:10.19678/j.issn.1000-3428.0061116.
[12] Xingyu Zhao & Shifei Ding. (2018). Review of deep reinforcement learning study. Computer Science(07),1-6.
[13] Xia Lei & Xionglin Luo. Review of Deep learning's explainability study. Computer Application.
[14] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need .