# A Study on the improvement of chinese automatic speech recognition accuracy using a lexicon

1st Min-Jeong Gu
*University of Science and Technology*
Daejeon, Republic of Korea
gmj1130203@gmail.com

2nd Shin-Gak KANG
ETRI
Daejeon, Republic of Korea
sgkang@etri.re.kr

***Abstract*—In this paper, in order to improve the error rate that occurs when Automatic Speech Recognition (ASR), one of the technologies widely applied in the field of speech recognition, is applied to Chinese, the study results on how to improve the Chinese recognition error rate by proposing a model to be added are described. As a result of testing by applying the xlsr-53 data set based on the Wav2vec2.0 model, it was confirmed that the Chinese recognition rate was improved.**

***Keywords—ASR Model, Word lexicon, Chinese speech recognition, Wav2vec 2.0***

## I. INTRODUCTION

With the development and dissemination of speech recognition technology, much attention is paid to improving the accuracy of speech recognition results. Various methods have been proposed to improve the accuracy of speech recognition, but there are still errors that need to be improved, such as homonyms and misrecognition, in order to be introduced into an expert system. A recently widely used technology in the field of speech recognition is automatic speech recognition (ASR). ASR is a technology that automatically executes the process of receiving voice input and outputting it as text.

In the past, experts directly adjust each parameter of the acoustic model (AM), the lexicon (Lexicon), and the language model (LM) using HMM (Hidden Markov Model) models and perform voice recognition. come. However, deep learning technology, which has recently been attracting attention in speech recognition, has enabled end-to-end learning that learns a target result (output) from data (input) without a separate intermediary [3].

Despite the advantage of being able to output a model that optimizes all parameters from start to finish just by putting it into a deep learning model, this method also has limitations. This method showed better performance when a large amount of data was given as input data, especially when labeled data was given as input data. However, it required a lot of labor, time, and work to create a labeled data set, which became a very difficult problem for researchers.

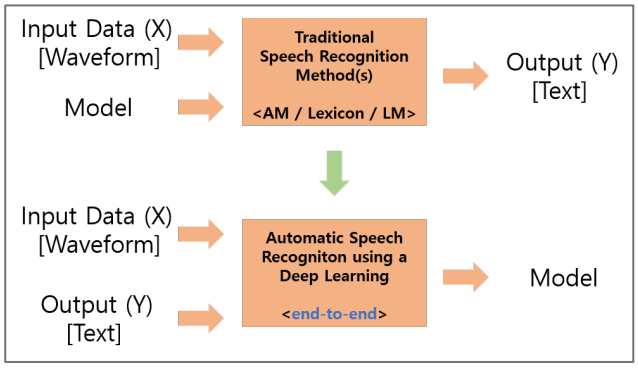


Fig. 1. Comparing with Traditional Method and Deep Learning Method

To solve the problem that requires a lot of labeled data, Facebook proposed a new voice recognition model, Wav2vec 2.0, in June 2020. The Wav2vec 2.0 model, which uses self-supervised learning and shows high accuracy even with a small amount of label data, is currently being actively used in the field of automatic speech recognition (ASR). It is mainly used in the form of fine-tuning the pre-trained wav2vec 2.0 model in the subject you want to apply.

This paper describes the research results on the improvement of the error rate that occurs when automatic speech recognition (ASR: Automatic Speech Recognition) is applied to Chinese. Basically, the wav2vec 2.0 model was used, and a method to improve the Chinese recognition error rate was proposed by adding Pinyin, the Chinese pronunciation symbol, to the Word Lexicon. And by applying the xlsr-53 data set, an experiment was performed to check whether the Chinese recognition rate was improved, and the experimental results were analyzed.

## II. OVERVIEW OF SPEECH RECOGNITION TECHNOLOGY

### *A. Before emergency of the Deep learning*

Traditional speech recognition technology has been developed based on statistics, and a representative model is the HMM-based model of Figure 2 using the Hidden Markov Model.

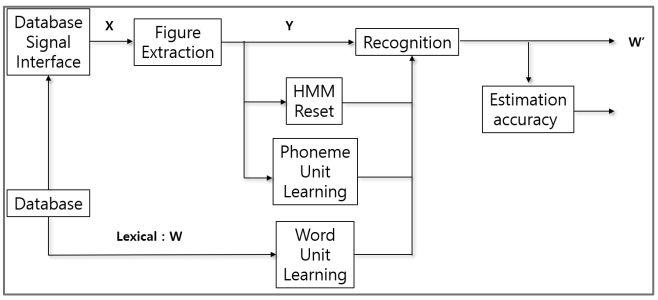


Fig. 2. Speech Recognition system structure using HMM Model.[1]

In the past, the speech recognition system was divided into three main parts: an acoustic model (AM), a lexicon (Lexicon), and a language model (LM). Unlike other data, voice data may include a variety of information in one input. This is a very attractive part because it can include various data such as the speaker's speech characteristics, intonation, surrounding environment, and speech space as well as the speech content, but it has a big difference in that it is continuous unlike other data.

### *B. After emergency of the Deep learning*

However, in modern times, a lot of data has been generated due to the development of the Internet and various electronic devices, and it was expected that voice patterns could be easily found based on a large amount of data. At that time, the deep learning technique advantageous for pattern matching became a big issue at the time, and we will try to apply it to the field of automatic speech recognition. Through several data preprocessing processes, it was possible to visualize the voice data with the characteristics of the waveform data and analyze the pattern. After applying the deep learning technique in the previous step, which obtained the loss in each of the existing three steps, the three processes are combined to make into one loss.

The data preprocessing process for converting voice data into visual data that can find patterns was somewhat complicated, but in the end, the input voice data was Fourier transformed and expressed as a frequency spectrogram, which is then weighted to a low frequency band similar to human speech frequency. It is possible to image using the vector value converted to Mel-spectrogram.

## III. DEEP LEARNING BASED ASR MODEL

Recently, along with the development of the deep learning technology mentioned above, research to apply the deep learning technology to automatic speech recognition is being actively promoted. The representative ASR basic structure that takes an end-to-end format by integrating the existing three steps into a decoder by applying deep learning technology is the 'Encoder-Decoder' model shown in Figure 3.

### *A. Encoder – Decoder (Sequence to Sequence)*

The basic structure of the encoder-decoder model is shown in same as Figure 3 The encoder compresses and expresses the input voice data, and the decoder plays a role in converting the compressed voice data into text. The encoder receives data and compresses information into a single vector. This compressed vector data is called a context vector, and the decoder converts the context vector back into text data format. Whenever an encoder receives a value, it is stored in the context vector. When new data is input, it is accumulated in the context vector and input and stored. After going through this process, at the end, the context vector accumulates and stores all input data before. If the input value is excessively generated in the context vector, the encoder compresses the input information and converts it into a fixed-length vector. Finally, the encoder result is stored in a context vector named h.

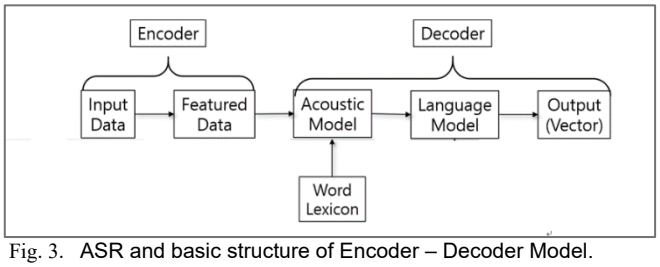


Fig. 3. ASR and basic structure of Encoder – Decoder Model.

The decoder converts the above context vector into an RNN-based language model (LM). Decoder is a value that comes out through the RNN-based model (usually LSTM) and the softmax function. (LSTM model - Affine layer - Softmax layer // Affine layer: This is a network that receives the hidden state as an input and outputs the number of classifications.)
However, the encoder-decoder structure of seq2seq has a limitation in that the performance deteriorates when the input data becomes long, as mentioned in the above encoder structure. If the encoder does not effectively compress the input data, it may not include all of the key information, resulting in a decrease in accuracy. Therefore, the attention mechanism was used to compensate for this problem when compressing the input data, and it shows a very improved result.

### *B. Self-supervised Learning*

In the case of speech recognition, in particular, the amount of labeled correct answer data was very small, which caused many difficulties. In the case of voice data, the acoustic data and the language model match well, so that the text similar to the correct answer data as much as possible matches the align. The self-supervised learning method has solved this problem, and the initial model applied to the ASR field is the CTC

model. However, although CTC was meaningful in its early attempts in the speech domain, it did not deal with speech recognition as a major topic.[5] Subsequently, an initial version of wav2vec was introduced, which was a model that specialized in applying CTC to voice [6]. In another method, the VQ wav2vec model [6), BERT's MLM pre-training was introduced into the voice by adding a quantization module. The model based on the existing CTC method receives voice data continuously, but the biggest difference is that the model receives it discretely.

Lastly, the Wav2vec 2.0 model [7] is a voice recognition specialized model that applied deep learning technology announced in 2016 by Facebook. The Acoustic Model stage, which implements the wavelength data unique to voice as a Mel-Spectrogram that can be recognized by humans with the end-to-end technique, and the Language Model stage, which manages the decoded text data as an output result It is a model implemented by combining . This model is similar to the VQ wav2vec model, but it is different from the existing method in that it supplements the quantization module and changes the structure to enable end-to-end learning. In the case of the Wav2vec2.0 model, which is based on self-supervised learning, it is a model that shows good learning performance even with a small amount of labeled correct answer data, unlike the models that previously required a large amount of labeled data. Labeled data is pre-trained, and unlabeled data is fine-tuned.

And in December 2020, Facebook announced XLSR-Wav2vec 2.0 Unsupervised Cross-lingual Representation Learning using a multilingual dataset called xlsr-53 on Facebook. [10] As a result of this study, the learning process was simplified by learning without using a language model (LM), but the error rate was increased. That is, in the case of the end-to-end learning method including the existing Wav2vec 2.0 model, unlike the traditional voice recognition, phoneme-mediated training and the Lexicon dictionary are omitted.

## IV. Problem Statement and Proposal

Various studies related to speech recognition have been actively carried out, but research on 'English' is mostly. English is the most used language in the world, and since the language system has a relatively simple structure of 26 alphabets, it has the advantage of having a structure that is easy to directly experiment with theoretical models. In particular, in the case of a model using deep learning, a large amount of data is required, and since it is easy to experiment with a relatively simple language system, English is overwhelmingly more common than other languages. Therefore, even if the published speech recognition model is applied to other languages such as Korean or Chinese and attempts to verify the recognition result, the data set is insufficient. In particular, in the case of Korean and Chinese, the number and amount of publicly available labeled data that anyone can use for model validation are small. Although the multilingual data set called xlsr-53 provided by Facebook is insufficient, it provides data that can conduct speech recognition research on Chinese to some extent.

The Wav2vec 2.0 model is recognized as a good model that shows high performance in speech recognition research using deep learning, and is currently a universally used model in the automatic speech recognition (ASR) field. If this is not sufficient, there is a limit to the study of speech recognition for the corresponding language. In other words, although a small amount of labeled data shows good performance, the more high-quality labeled data is used, the better the speech recognition accuracy is. Considering this environment, this study proposes a model that can improve the error rate even when using a small amount of dataset in Chinese speech recognition, and analyzes the speech recognition results according to the proposed model.

Many studies have already been published confirming that the performance of automatic speech recognition is improved when the vocabulary dictionary is well constructed in speech recognition [8]. However, there are few studies that have actually investigated the correlation between recognition rates using official phonetic symbols approved by the state.

In the case of Chinese, since it has quite a lot of characters, there are not only officially designated phonetic symbols, but also the official Chinese pronunciation added to the Word Lexicon for Chinese, which is a language designated in the form of the English alphabet. We analyzed how pinyin, the symbol, affects the accuracy of speech recognition.

In this study, as described above, the Wav2vec2.0 model [4], which is the most widely used among the models applying deep learning in the automatic speech recognition (ASR) field, was used. The step of referring the Lexicon dictionary to the existing Wav2vec2.0 model was added as a post-processing concept after the acoustic model (AM). suggested. In order to verify the performance of the proposed model, it is necessary to compare the performance with the existing model under the same conditions, so the error rate was compared and analyzed using the pre-trained wav2vec2.0.

## V. Experiment

### A. Experiment Environments

In this study, 'Fairseq', that is, the Wav2vec2.0 model announced by Facebook, which was published on github, an open source code community site, is used as an automatic speech recognition model. In addition, the data set used for the experiment used 50 hours of Chinese data from the xlsr-53 multilingual package for 53 languages provided on github. The Chinese data set used contains 7,176 sentences and 169,193 words.

As the development environment for this experiment, two NVIDIA GPU RTX Quadra 8000 GPUs and 64 CPU cores were used. Also, as a result of performing some experiments using the virtualized GPU resources provided by Google's CoLab environment, there was no particular difference from the experiments in the local environment.

### B. Establishment of experiment model

As shown in Figure 4, the model proposed in this paper to improve the speech recognition error rate is

based on the existing Wav2vec 2.0 model and presents an improved structure by adding Chinese phonetic symbols to the vocabulary dictionary. In addition to the vocabulary existing in the existing language model stage, as shown in Figure 6 below, In addition to the vocabulary existing in the existing language model stage, as shown in Figure 6 below, a total of 398 phonetic symbols that combine 23

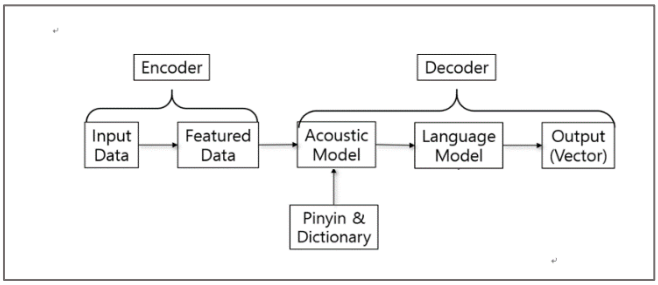


Fig. 4. Proposed Model – Add a pronounciation symbol 'Pinyin' put in Language word dictionary.

Chinese consonants and 35 vowels are included in the vocabulary dictionary. Chinese phonological rules can be broadly divided into falsification (變調), light tone (轻声) and 'Er-hwa' (儿化). In Chinese, each syllable is a unit with a meaning, so there is almost no phenomenon in which the sound changes as in Korean. (9) Therefore, it can be said that the method of putting all combinations following the alphabetical formula in the vocabulary is an effective method in reducing the error rate. An example of adding phonetic symbols to the vocabulary is shown in Figure 5.

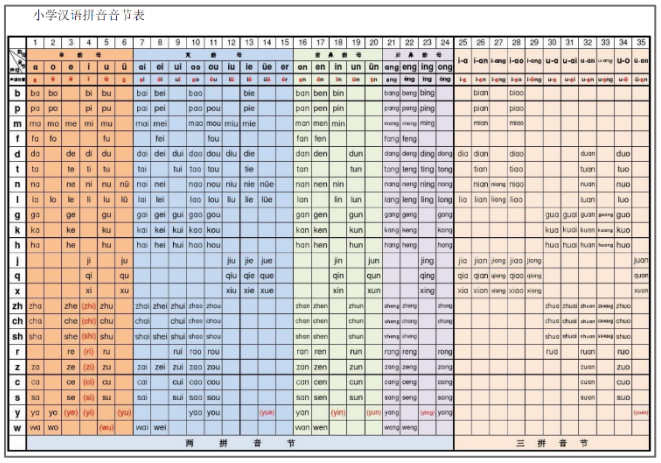


Fig. 5. Chinese pinyin Consonant vowel combination table [12]

Figure 5 shows all Chinese alphabets with phonetic symbols, and syllables that cannot be pronounced or are not used are not marked. In the case of this material, it has the name ' 小學漢語拼音音節表 ', and it is a pronunciation chart certified by the Ministry of Education, a national institution of China, and is actually a pronunciation symbol taught in the elementary education course.

In addition, considering the characteristic that tones are attached only to vowels as a characteristic of Chinese language itself, a combination table was created by inserting the numbers 1, 2, 3, 4, meaning 4 Chinese characters, after all vowels. Therefore, the entire combination was trained to refer to a total of 1,592 existing pronunciation possible combinations, 398 × 4 adults. Unlike Korean, which belongs to an agglutinative language, in which parts of speech are changed by adding an additional proposition, Chinese has the characteristic of language isolate that parts of speech change depending on location. In other words, there is less articulation change, which is a change in pronunciation, compared to English or Korean.

Data pre-processing and feature extracted vector values are obtained from the waveform form, which is the first raw data, and the vector value with the highest probability can be finally output as text by referring to Lexicon.

| | | | | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | ba | 17 | pa | 34 | ma | 53 | fa | 62 | da | 85 | ta | 104 | na | 128 | la | | | | | | | | |
| 2 | bo | 18 | po | 35 | mo | 54 | fo | 63 | de | 86 | te | 105 | ne | 129 | lo | | | | | | | | |
| 3 | bi | 19 | pi | 36 | me | 55 | fu | 64 | di | 87 | ti | 106 | ni | 130 | le | | | | | | | | |
| 4 | bu | 20 | pu | 37 | mi | 56 | fei | 65 | du | 88 | tu | 107 | nu | 131 | li | | | | | | | | |
| 5 | bai | 21 | pai | 38 | mu | 57 | fou | 66 | dai | 89 | tai | 108 | nu: | 132 | lu | | | | | | | | |
| 6 | bei | 22 | pei | 39 | mai | 58 | fan | 67 | dei | 90 | tui | 109 | nai | 133 | lu: | | | | | | | | |
| 7 | bao | 23 | pao | 40 | mei | 59 | fen | 68 | dui | 91 | tao | 110 | nei | 134 | lai | | | | | | | | |
| 8 | bie | 24 | pou | 41 | mao | 60 | fang | 69 | dao | 92 | tou | 111 | nao | 135 | lei | | | | | | | | |
| 9 | ban | 25 | pie | 42 | mou | 61 | feng | 70 | dou | 93 | tie | 112 | nou | 136 | lao | | | | | | | | |
| 10 | ben | 26 | pan | 43 | miu | | | 71 | diu | 94 | tan | 113 | nu | 137 | lou | | | | | | | | |
| 11 | bin | 27 | pen | 44 | mie | | | 72 | die | 95 | tun | 114 | nie | 138 | liu | | | | | | | | |
| 12 | bang | 28 | pin | 45 | man | | | 73 | dan | 96 | tang | 115 | nu:e | 139 | lei | | | | | | | | |
| 13 | beng | 29 | pang | 46 | men | | | 74 | den | 97 | teng | 116 | nan | 140 | lu:e | | | | | | | | |
| 14 | bing | 30 | peng | 47 | min | | | 75 | dun | 98 | ting | 117 | nen | 141 | lan | | | | | | | | |
| 15 | bian | 31 | ping | 48 | mang | | | 76 | dang | 99 | tong | 118 | nin | 142 | lin | | | | | | | | |
| 16 | biao | 32 | pian | 49 | meng | | | 77 | deng | 100 | tian | 119 | nang | 143 | lun | | | | | | | | |
| | | 33 | piao | 50 | ming | | | 78 | ding | 101 | tiao | 120 | neng | 144 | lang | | | | | | | | |
| | | | | 51 | mian | | | 79 | dong | 102 | tuan | 121 | ning | 145 | leng | | | | | | | | |
| | | | | 52 | miao | | | 80 | dia | 103 | tuo | 122 | nong | 146 | ling | | | | | | | | |
| | | | | | | | | 81 | dian | | | 123 | nian | 147 | long | | | | | | | | |
| | | | | | | | | 82 | diao | | | 124 | niang | 148 | lia | | | | | | | | |
| | | | | | | | | 83 | duan | | | 125 | niao | 149 | lian | | | | | | | | |
| | | | | | | | | 84 | duo | | | 126 | nuan | 150 | liang | | | | | | | | |
| | | | | | | | | | | | | 127 | nuo | 151 | liao | | | | | | | | |
| | | | | | | | | | | | | | | 152 | luan | | | | | | | | |
| | | | | | | | | | | | | | | 153 | luo | | | | | | | | |

Fig. 6. Example of Chinese pinyin Lexicon.

## C. *Experiment result*

In the previous study [10], when the wav2vec2.0 model was run for 1 hour using the xlsr-53 data set, the error rate (PER: Phoneme Error Rate) was reported to be 18.3%. In addition, as a result of running the existing model on its own using the given experimental environment, the PER was measured to be 18.8%, and the sub error rate generated by the synthesis was measured to be 12%. The result of running the model in which Pinyin, the Chinese phonetic symbol, is added to the lexical dictionary proposed in this paper in the same environment is shown in Table 1.

Table 1. Table of adding pronounciation symbol in dictionary result

| SPKR | # Snt | # Wrd | Corr | Sub | Del | Ins | Err | S.Err |
|---|---|---|---|---|---|---|---|---|
| none | 7176 | 169193 | 83.8 | 8.2 | 3.8 | 2.6 | 14.6 | 94.6 |
| Sum/Avg | 7176 | 169193 | 83.8 | 8.2 | 3.8 | 2.6 | 16.6 | 94.6 |
| Mean | 7176 | 169193 | 83.8 | 8.2 | 3.8 | 2.6 | 14.6 | 94.6 |
| S.D. | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Median | 7176 | 169193 | 83.8 | 8.2 | 3.8 | 2.6 | 14.6 | 94.6 |

As can be seen from the results of this experiment, the recognition error rate was measured to be 14.6%, confirming that the error rate was reduced. Also, the rate of sub error was reduced to 8.2%. This result means that the error rate due to misrecognition can be reduced when automatic speech recognition is executed by including phonetic symbols (that is, pinyin) in the lexical dictionary.

## VI. CONCLUSION

In this paper, we conducted a study on how to improve the error rate of Chinese speech recognition using the Wav2vec 2.0 model, which is a representative automatic speech recognition model published by Facebook. Based on the Wav2vec 2.0 model, a kind of post-processing was performed so that the lexicon can be additionally referenced after the acoustic model (AM).

As a result of the experiment, adding Chinese pronunciat
-ion information to the lexicon of the existing model reduced the speech recognition error rate. was meaningful in the study. In this study, the xlsr-53 data set was used to analyze the improvement of the error rate of Chinese speech recognition, but the provided data set is relatively scarce compared to English, etc,. so it is thought that there is a limit to improving the error rate. In the future, it is expected that more improved results can be obtained if a richer data set is secured and the proposed model of this study is applied to training.

## ACKNOWLEDGMENT

1. This work was supported by Electronics and Telecommunications Research Institute(ETRI) grant funded by the Korean government [21YR 1200, Enhancement of ETRI Open Source Governance and Supporting Open R&D Activity (2021.09.01.-22.04.30)

2. If you intend to utilize the contents of this paper, you must disclose that the research was funded by Electronics and Telecommunications Research Institute (ETRI)

## REFERENCES

[1] Park, Yoo-hyun. "Quantitative Analysis of Gartner's." Journal of the Korea Institute of Information and Communication Engineering 22.8 (2018): 1041-1048.

[2] Lee, Suji, et al. "Korean speech recognition using deep learning." The Korean Journal of Applied Statistics 32.2 (2019): 213-227.

[3] Davis, Ken H., R. Biddulph, and Stephen Balashek. "Automatic recognition of spoken digits." The Journal of the Acoustical Society of America 24.6 (1952): 637-642.

[4] Lee, Gun-sang, “Speech Recognition”, Hanyang Univ. , 2001, p.28, Figure 2.1

[5] Oord, Aaron van den, Yazhe Li, and Oriol Vinyals. "Representation learning with contrastive predictive coding."arXiv:1807.03748 (2018).

[6] Schneider, Steffen, et al. "wav2vec: Unsupervised pre-training for speech recognition." arXiv preprint arXiv:1904.05862 (2019).

[7] Baevski, Alexei, Steffen Schneider, and Michael Auli. "vq-wav2vec: Self-supervised learning of discrete speech representations." arXiv preprint arXiv:1910.05453 (2019).

[8] Baevski, Alexei, et al. "wav2vec 2.0: A framework for self-supervised learning of speech representations." arXiv preprint arXiv:2006.11477 (2020).

[9] Jang, Tae-Yeoub, “Implementation of a non-native pronunciation dictionary for automatic recognition of utterances by Korean learners of English”, (Journal of humanities) 56 pp.99~122, (2006).

[10] Cho, Ara, " Finding ways to educate Chinese learners on Korean pronunciation(중국인학습자의 한국어 발음교육방안 모색)." , Korean Language in China (2), (2019): 50-63.

[11] Conneau, Alexis, et al. "Unsupervised cross-lingual representation learning for speech recognition." arXiv preprint arXiv:2006.13979 (2020)

[12] Vaswani, Ashish, et al. "Attention is all you need." Advances in neural information processing systems. 2017

[13] https://www.51wendang.com/doc/dd97ebf4e656d3a950a1c397e408710308690493