Journal of Electronics and Information Science (2024)
Clausius Scientific Press, Canada

DOI: 10.23977/jeis.2024.090318
ISSN 2371-9524 Vol. 9 Num. 3

Study on Speech Recognition Optimization of Cross-modal Attention Mechanism in Low Resource Scenarios

Zhuoxi Li[*]

School of Computing, Guangdong Neusoft University, Foshan, China
15122597559@163.com
[*]*Corresponding author*

Keywords: Cross-modal attention mechanism; Low resource scenario; Speech recognition; Multimodal information fusion

Abstract: With the rapid development of artificial intelligence technology, speech recognition technology has become one of the important interfaces of human-computer interaction, and its accuracy and robustness are crucial to user experience. However, in low-resource scenarios, such as noise interference, dialect accents, and limited labeled data, the performance of speech recognition systems often deteriorates significantly. To solve this problem, a speech recognition optimization method based on cross-modal attention mechanism is proposed in this paper. As a new technique, cross-modal attention mechanism provides a new idea for speech recognition in low resource scenarios. By integrating information from different modes (such as vision, text, etc.), the mechanism makes use of the complementarity between them to enhance the recognition ability of the model. In speech recognition tasks, audio signals and visual information associated with them (such as lip movements, gestures, etc.) are often strongly correlated. Through the cross-modal attention mechanism, the model can pay more attention to the visual features closely related to the speech content, so as to achieve accurate recognition of audio signals. This paper first introduces speech recognition technology and its challenges in low resource scenarios, and discusses its application strategy in low resource speech recognition in detail by analyzing the basic principle of cross-modal attention mechanism. By introducing an attention mechanism, the neural network can automatically learn and selectively focus on important information in the input, thereby improving the performance and generalization ability of the model.

1. Introduction

Speech recognition is an important research direction in the field of artificial intelligence, which aims to convert human speech signals into text information, so as to realize effective interaction between humans and machines. In recent years, with the continuous progress of deep learning technology, the performance of speech recognition systems has been significantly improved. However, in practical applications, especially in low-resource scenarios, speech recognition still faces many challenges. In this case, the traditional training methods based on a large number of labeled data are difficult to achieve ideal results, resulting in a significant decline in the accuracy of speech recognition systems. In order to deal with the challenge of speech recognition in low-resource scenarios, researchers have proposed a variety of optimization strategies. These strategies mainly focus on data enhancement, model transfer learning, semi-supervised learning and unsupervised learning. While these methods alleviate the problem of data scarcity to some extent, they often rely on additional resources or assumptions and are still difficult to achieve satisfactory results in some extreme cases. Therefore, it is of great practical significance to explore more efficient and robust speech recognition optimization methods.

2. Speech recognition Overview

2.1 Basic Principles of speech recognition

Speech recognition is a technology that converts human speech signals into text messages that machines can understand. Its basic principle includes the steps of speech signal preprocessing, feature extraction, acoustic model modeling and language model decoding.

First of all, the speech signal needs to be preprocessed, including denoising, framing, windowing and other operations. Denoising is to remove background noise in speech signal and improve signal quality. Framing is the segmentation of continuous speech signals into several short-time frames for subsequent processing. Adding Windows is to reduce the problem of spectrum leakage caused by signal truncation. Next, feature extraction is carried out. Feature extraction is to extract the feature information useful for speech recognition from the pre-processed speech signal. Commonly used features include Maier frequency cepstrum coefficient (MFCC), linear predictive cepstrum coefficient (LPCC) and so on. These features can reflect the spectrum characteristics and fundamental frequency information of speech signals, and are very important for the subsequent acoustic modeling. Acoustic models are then used to model the extracted features. Acoustic model is one of the core components of speech recognition system, which is used to describe the mapping relationship between speech signal and text. Commonly used acoustic models include Hidden Markov model (HMM), deep neural network (DNN), etc. [1]. With the development of deep learning techniques, deep neural networks have gained significant advantages in acoustic modeling. Finally, the language model is decoded. The language model is used to describe the probability distribution between text sequences, which can help the acoustic model select the most syntactic and semantic text sequences from the candidate texts. Commonly used language models include N-gram (N-gram), recurrent neural network language model (RNNLM), etc. Based on the output of acoustic model and language model, the decoder uses dynamic programming or heuristic search to find the optimal text sequence.

2.2 Voice Recognition Challenges in low-resource Scenarios

(1) Annotation data is scarce

The scarcity of annotation data is one of the biggest challenges faced by speech recognition in low-resource scenarios. Traditional speech recognition systems rely on a large amount of annotation data for training to learn the mapping relationship between speech signals and text[2]. However, in low-resource scenarios, annotated data is often very limited or even unavailable. This results in inadequate model training and difficulty in learning effective feature representation and mapping relations, thus affecting the accuracy of recognition.

(2) Serious noise interference

Noise interference is another important challenge for speech recognition in low-resource scenarios. In practical applications, speech signals are often disturbed by various noises, such as background noise, echo, noise and so on. These noises can interfere with the clarity and accuracy of the speech signal, thus affecting the performance of recognition. In low-resource scenarios, due to the scarcity of labeled data, the robustness of the model to noise is poor, and it is easy to be disturbed by noise, leading to identification errors.

(3) Poor voice quality

Due to equipment limitations, environmental interference and other reasons, the collected speech signals often have poor quality problems, such as fuzzy sound, uneven speed, and unclear pronunciation. These problems will cause the feature representation of the speech signal to be inaccurate, thus affecting the accuracy of recognition. In low-resource scenarios, due to the scarcity of annotation data, the model has poor adaptability to speech quality and is easily affected by poor speech quality, resulting in poor recognition effect[3].

In response to the above challenges, researchers have proposed a variety of methods to solve the speech recognition problem in low-resource scenarios. However, these methods still have some limitations. For example, although the data enhancement method can increase the diversity of training data, it cannot fundamentally solve the problem of the scarcity of annotation data. Although transfer learning methods can use the knowledge of other tasks to improve the performance of speech recognition, it is difficult to guarantee the correlation between different tasks. Although the semi-supervised learning method can be trained using unlabeled data, its performance depends on the quantity and quality of labeled data. Therefore, it is necessary to explore new methods to solve the speech recognition problem in low resource scenarios.

3. Overview of cross-modal attention mechanism

3.1 Basic principle of cross-modal attention mechanism

The cross-modal attention mechanism is a deep learning technique whose goal is to achieve information fusion between different modes. This mechanism allows neural networks to focus on and integrate relevant information from different modes when processing multimodal data. Its core idea is to imitate the way human visual and cognitive systems selectively pay attention to information. By introducing attention mechanisms, neural networks can automatically learn and selectively pay attention to important information in the input, thereby improving the performance and generalization ability of the model[4].

The cross-modal attention mechanism introduces additional input sequences on top of the self-attention mechanism to fuse information from multiple sources for more accurate modeling. Specifically, it computes the attention matrix and applies it to the sequence of values to generate the final output sequence by asymmetrically combining two embedded sequences of the same dimension, one of which is used as a Query input and the other as a Key and Value input. This mechanism allows the neural network to automatically focus on and integrate the relevant information from different modes to improve the performance of the model when processing multi-modal data.

3.2 Implementation of cross-modal attention mechanism

The implementation of cross-modal attention mechanism usually includes the following steps: First, feature representations need to be extracted from inputs of different modes, which can be low-level sensory features (such as the spectral features of audio signals, pixel values of images, etc.) or high-level semantic features (such as the word vector representation of text, the semantic segmentation results of images, etc.)[5]. Then, the correlation calculation is carried out to evaluate the correlation or complementarity between different modes by calculating the dot product and cosine similarity between feature vectors or by using deep learning models (such as neural networks). Then, according to the calculated correlation or complementarity, differentiable operations such as softmax function are used to assign different weights to the features of different modes, which reflect their importance in the fusion process. Finally, the features of different modes are weighted and summed according to the assigned weights to obtain a feature representation that integrates complementary information from different modes, which can be used for subsequent speech recognition tasks (see Table 1).

Table 1: Shows the implementation process of the cross-modal attention mechanism

Implementation phase	Description	Example feature
Feature extraction	Feature representations are extracted from input data of different modes	The spectral characteristics of audio signal, the original pixel value of image, the word vector representation of text, and the semantic segmentation results of image
Correlation calculation	Calculate the correlation or complementarity between feature vectors	/
Weight distribution	According to the degree of correlation or complementarity, the features of different modes are weighted	/
Weighted sum	The features of each mode are weighted and summed according to the assigned weights	/
Result output	Generate feature representations that incorporate complementary information from multiple modes	Synthetic feature representation

4. Application of cross-modal attention mechanism to speech recognition in low-resource scenarios

4.1 Audio-Visual integration

In audio-visual integration, the cross-modal attention mechanism is applied to integrate audio signals with rich visual information such as lip movements and gestures. This technological innovation has shown broad application prospects in video speech recognition, remote conference communication optimization, deaf-mute assisted communication system and other scenarios[6]. In the process of audio feature extraction, a series of key audio features such as spectral features and Maier frequency cepstrum coefficient (MFCC) are extracted from audio signals through precise algorithms. These features are like the acoustic fingerprint of language, which can accurately reflect the acoustic characteristics of speech and lay a solid foundation for subsequent speech recognition. In terms of visual feature extraction, with the help of advanced image processing technology, the subtle movements of lips and the dynamic changes of gestures are meticulously captured from video frames. Like the visual mapping of speech, these visual features effectively capture visual cues closely related to speech content and provide valuable supplements for the deep integration of cross-modal information[7]. The extracted audio features and visual features are then input into the cross-modal attention mechanism module, which dynamically assigns reasonable weights to the features of different modes by calculating the correlation or complementarity between audio and visual features. Finally, these weights are applied to the weighted summation of audio and visual features to obtain the fused feature representation. This fused feature representation contains both the acoustic information of the audio signal and the visual information related to the speech, thus enhancing the robustness and accuracy of the speech recognition system.

4.2 Audio-Text Fusion

In the field of audio-text fusion, cross-modal attention mechanism has been applied to integrate audio signals and text information, which shows great potential application value in tasks such as speech translation, automatic title generation and speech search. In the process of audio feature extraction, key audio features such as spectral features and Maier frequency cepstrum coefficient (MFCC) are first extracted from audio signals. As acoustic fingerprints of speech, these features can accurately reflect the physical characteristics of speech and provide a solid foundation for subsequent fusion processing[8]. In terms of text feature extraction, advanced natural language processing technology is used to convert text information into word vectors, sentence vectors and even higher-level semantic representations. These text features are like semantic passwords of languages, which can profoundly capture the internal meaning and context information of languages and provide rich semantic support for cross-modal integration. Then, these extracted audio features and text features are input into the cross-modal attention mechanism module, which accurately measures the correlation between audio and text features through a complex calculation process, and dynamically assigns reasonable weights to the features of different modes, which reflects the importance of audio and text information in the fusion process. Finally, the weighted summation of these weighted audio and text features is carried out to obtain the fused feature representation. This feature represents the combination of acoustic information of audio signal and semantic information of text, which helps to improve the accuracy of speech recognition.

4.3 Multi-mode Fusion Strategy

In practical application, different multi-mode fusion strategies can be adopted according to specific scenarios and requirements. For example, in the pursuit of comprehensive and accurate information fusion, audio - visual fusion and audio - text fusion can be organically combined to achieve a more comprehensive and in-depth multi-modal information fusion through the integration of audio signals with visual information such as lip movements and gestures, as well as the supplement of text semantic information[9-10]. Further, in order to fully tap and utilize the potential of multi-modal information, information of other modes can be introduced, such as non-verbal information such as gestures and expressions, which often contain rich emotions and intentions, and can further improve the intelligence level and user experience of speech recognition systems (see Table 2).

In order to integrate multi-modal information effectively, it is necessary to design a reasonable network structure and algorithm to support the efficient operation of the cross-modal attention mechanism. For example, multi-head attention mechanism can be used to improve the efficiency and accuracy of information fusion by processing the correlation between different modes in parallel. At the same time, with the help of deep learning models such as convolutional neural network (CNN) or recurrent neural network (RNN), more advanced and abstract feature representations can be extracted, providing a richer and more accurate information basis for cross-modal fusion. Graph neural network (GNN), as a new network structure, also shows great potential in modeling complex relationships between different modes, and provides new ideas and methods for exploring and optimizing multi-mode fusion strategies.

Table 2: Multi-mode fusion strategy table

Fusion strategy	Application scenario	Advantage
Audio-visual fusion	Video speech recognition, deaf-mute assisted communication system	Combine lip movements and gestures to improve recognition accuracy; Enhance user experience
Audio-text fusion	Voice translation, automatic title generation, voice search	Combined with text semantics, improve the accuracy of speech recognition; Multi-language processing is supported
Audio-visual-text fusion	Remote conference communication optimization, intelligent customer service system	Comprehensively integrate a variety of information to improve communication efficiency and intelligence
Audio-posture fusion	Emotion recognition, human-computer interaction	Introduce gesture information to enhance emotional understanding and interactive experience
Audio-expression fusion	Sentiment analysis, intelligent education	Using expression information to improve the accuracy of emotion recognition; Enhance educational interaction

5. Conclusion

In this paper, a speech recognition method based on cross-modal attention mechanism is proposed to solve the speech recognition problem in low resource scenarios. By integrating audio signals with visual, text and other modal information, the cross-modal attention mechanism is used to automatically learn and selectively focus on the important information in the input. In low-resource scenarios, this method can effectively improve the accuracy of speech recognition. The cross-modal attention mechanism not only enhances the robustness of the model to noise, but also improves the adaptability to speech quality, thus showing great potential and application value in practical applications. In the future, we should continue to deepen the research on the mechanism of cross-modal attention and explore its application in more scenarios.

References

[1] Mao J,Shi H,Li X .Research on multimodal hate speech detection based on self-attention mechanism feature fusion[J].The Journal of Supercomputing,2024,81(1):28-29.

[2] Abderrazzaq M,David R,Pejman R .Attention-Based Fusion of Ultrashort Voice Utterances and Depth Videos for Multimodal Person Identification[J].Sensors (Basel, Switzerland),2023,23(13)

[3] Lin F,Yao L L, Lan S L, et al.Multimodal speech emotion recognition based on multi-scale MFCCs and multi-view attention mechanism[J].Multimedia Tools and Applications,2023,82(19):28917-28935.

[4] Yang L, Haoqin S, Wenbo G, et al.Multi-modal speech emotion recognition using self-attention mechanism and multi-scale fusion framework[J].Speech Communication,2022,139.

[5] Mehra S, Ranga V, Agarwal R .Dhivehi Speech Recognition: A Multimodal Approach for Dhivehi Language in

Resource-Constrained Settings[J].Circuits, Systems, and Signal Processing,2024,(prepublish):1-21.

[6] Zhou K, Yang J, Loy C C, et al. Learning to prompt for vision-language models[J]. International Journal of Computer Vision, 2022, 130(9):2337-2348.

[7] Li J, Selvaraju R, Gotmare A, et al. Align before fuse: Vision and language representation learning with momentum distillation[J].Advances in neural information processing systems,2021,34:9694-9705.

[8] Brown T, Mann B, Ryder N, et al. Language models are few-shot learners[J]. Advances in neural information processing systems, 2020, 33: 1877-1901.

[9] Lu J, Batra D, Parikh D, et al. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks [J].Advances in neural information processing systems, 2019, 32.

[10] Frome A, Corrado G S, Shlens J, et al. Devise: A deep visual-semantic embedding model[J].Advances in neural information processing systems,2013, 26.