Transactions on Computer Science and Intelligent Systems Research
ISSN: 2960-1800, eISSN: 2960-2238 | Volume 5
AIDML 2024

A Comparative Analysis of LSTM and Transformer-based Automatic Speech Recognition Techniques

Ruijing Zhang *

Department of Information Engineering, Chinese University of Hong Kong, Hong Kong, 999077, China

* Corresponding Author Email: 1155191443@link.cuhk.edu.hk

Abstract. Automatic Speech Recognition (ASR) is a technology that leverages artificial intelligence to convert spoken language into written text. It utilizes machine learning algorithms, specifically deep learning models, to analyze audio signals and extract linguistic features. This technology has revolutionized the way that people interact with voice-enabled devices, enabling efficient and accurate transcription of human speech in various applications, including voice assistants, captioning, and transcription services. Among previous works for ASR, Long Short-Term Memory (LSTM) networks and Transformer-based methods are typical solutions towards effective ASR. In this paper, the author focuses on an in-depth exploration of the progression and comparative analysis of deep learning innovations within the ASR domain. This work starts with a foundational historical perspective, mapping the evolution from pioneering ASR systems to the current benchmarks: LSTM networks and Transformer-based models. The study meticulously evaluates these technologies, dissecting their strengths, weaknesses, and the potential they hold for future advancements in ASR.

Keywords: Speech recognition; deep learning; long short-term memory, transformer.

1. Introduction

Automatic Speech Recognition (ASR) is a technique applying the machine learning and Artificial Intelligence (AI) to help the computer program to understand the human language and transform it into a coding command.

In modern lives, it takes part in multiple regions such as transcription between speech and text, the virtual assistant like Siri, the real-time barrage used in TikTok and YouTube. As main methods nowadays used in ASR, Big Data and AI have their own advantage. The Big Data collects the needed data, achieves the pre-processing and does the modeling, providing the raw material for the AI. While AI can achieve the machine learning and the subsequent Natural Language Processing (NLP).

The ASR goes through a long way, from the Hidden Markov Model [1], the Large Vocabulary Continuous Speech Recognition [2], and the AI technique [3]. Among them, the Long Short-Term Memory (LSTM) model help to solve the problem of vanishing gradient in Recurrent Neural Networks (RNNs) as one of the deep learning [4]. At the same time, the initially used in machine translation Transformer architecture has taken the advantage of the computation effectiveness and parallelism over the traditional Convolutional Neural Network (CNN) model. It is used in promoting the contextual information capture in ASR. Both the improvements make contribution to the adaptation to the complex real-life application, like the localism. The article will introduce the Transformer-based and LSTM-based ASR and analyses the advantage and disadvantage of them.

2. LSTM-based ASR

Deep learning, with its inherent ability to automatically detect features, handle vast quantities of data, and continuously learn from new inputs, holds a pivotal position in the field of Automatic Speech Recognition (ASR) [5]. Among the various deep learning models, the LSTM stands out as a unique architecture that effectively addresses the vanishing gradient problem encountered in RNNs.

The LSTM comprises three fundamental components, each playing a crucial role in its functionality. Firstly, the forget gate serves to discard irrelevant information, ensuring that only pertinent data is retained. Secondly, the input gate is responsible for updating the network with necessary new information. Finally, the output gate functions as a conduit, transporting the processed information to subsequent layers or for external utilization.

By integrating these three gates, the LSTM is able to capture long-term dependencies in sequential data, making it particularly suitable for tasks such as speech recognition, where the context of previous and subsequent words is crucial for accurate interpretation. Its adaptability and effectiveness in handling complex temporal patterns have contributed to its widespread adoption in various ASR systems.

2.1. Representative Works

Previous work has tackled speech enhancement using LSTM [6]. The enhancement is achieved by predicting time-frequency masks from the magnitude spectrum of a noisy signal. Given the estimated mask, the time frame, and the magnitude, one can represent a noise signal. The method for separation is straightforward, aiming to minimize the Signal-to-Noise Ratio (SNR). Theoretically, for different layers and phases, the article proposes discriminatively training an LSTM-DRNN differently to perform the initial approximation. The integration of ASR information further aids in separating the sentence. Finally, a multi-channel extension is employed to prevent overfitting, rather than relying on a single-channel approach. In the experiments, Signal-to-Distortion Ratio (SDR) and Word Error Rate (WER) were set as evaluation metrics to assess speech separation and ASR performance. The results demonstrate that the relationship between SNR and WER contradicts previous research, indicating that speech separation using recurrent neural networks can be effectively utilized as a front-end to enhance the noise robustness of state-of-the-art acoustic models for ASR [7].

Another paper achieves feature enhancement by using multiple hidden layers for extracting progressively higher-level representations [8]. The architecture preserving the temporal context of RNN and employing supervised learning for non-linear mappings from noisy to clean speech features. In the experiment part, training the model with end-pointed speech segment is also a strong characteristic of it. Going through the Network training the map the features of the noisy training set of the above-mentioned corpus to a noise-free training set and baseline networks to verify the effectiveness of the feature enhancement and finally getting the ASR features, the paper set up a baseline model and input with feature transformations. In conclusion, the paper attends importance to frame-to-frame correspondences between distorted and clean training features, less dependency on the input and more on the pre-training part which will definitely increase its performance during work time.

2.2. Summary

LSTM-based ASR has several advantages:

Firstly, it contains long-term contextual understanding. Using its gating mechanisms, LSTMs excel at capturing long-term dependencies, making them effective for speech recognition tasks that really matters

Secondly, it has strong adaptability. LSTMs can be tuned for various aspects of speech recognition, ranging from noise reduction and feature enhancement to complex linguistic structures.

However, there are also several limitations.

Firstly, it has high computational complexity. Due to their complex architecture, LSTMs require significant computational resources, which can be a challenge if the input is not enough

Secondly, it has strong training difficulties. LSTMs are relatively harder to train compared to simpler neural network models, requiring more optimization to prevent issues like overfitting.

Overall, the LSTM bridges the gap for coherent integration. The incorporation of LSTM models into ASR has pushed the boundaries of understanding human speech. The capability of LSTMs to deal with the temporal variance and context of speech makes them invaluable to the progress of ASR technology. However, as with any technology, they come with their set of challenges, primarily related to computational demands and complexity in training.

The future direction might involve hybrid models that combine the strength of LSTMs in handling sequential data with the computational efficiency of other neural network architectures. Such integrations could address the current limitations, offering more robust, efficient, and adaptable ASR systems.

In summary, while LSTMs provide a solid foundation for tackling the complex requirements of ASR, continuous exploration and innovation in deep learning are essential for overcoming the inherent limitations of LSTM models. By focusing on enhancing the efficiency, adaptability, and ease of training, the next generation of ASR systems can achieve even higher levels of accuracy and reliability.

3. Transformer-based ASR

The Transformer is a revolutionary architecture in the field of deep learning, particularly for tasks involving sequence data. Its key innovation lies in its ability to capture long-range dependencies in sequences effectively through the use of self-attention mechanisms. Unlike traditional models, the Transformer doesn't process the input sequence in a strictly sequential manner. Instead, it employs a multi-head attention mechanism that allows each position in the sequence to attend to (or focus on) other positions, regardless of their distance [9].

3.1. Representative Works

The Speech-Transformer combines seq2seq and transformer [10]. It introduces a transformative no-recurrence, encoder-decoder framework in speech recognition, diverging from conventional sequential models through the adoption of multi-head attention and position-wise feed-forward networks. This innovative model encodes speech features into a hidden representation, which the decoder then utilizes to sequentially generate character outputs. Key features include Scaled Dot-Product Attention, enhancing input sensitivity through scaled weighting, Multi-Head Attention, which captures diverse data representations simultaneously, and a Position-Wise Feed-Forward Network, enabling deeper feature processing. The model's architecture, combining convolutional networks to process spectrograms and transformer blocks for encoding and decoding, introduces a novel 2D-Attention mechanism. This mechanism is specifically designed to capture the complex temporal and spectral dynamics inherent in speech, significantly enhancing the model's ability to accurately convert speech into text by attending to both time and frequency dimensions. The Speech-Transformer thus offers an advanced, efficient solution for automatic speech recognition by leveraging transformer architecture's unique capabilities to improve performance and reduce computational demands. In conclusion, the speech-transformer model converged with considerably small training costs and achieved competitive performance, which shows the efficiency and effectiveness of the Speech-Transformer. This model could be applied to several tasks. A typical example is in virtual voice assistant, due to the wide spread of it in real world that the request for quick response, accuracy and different language. The pre-training models acquire more contextual content but enjoy a higher accuracy and speed.

The ASR tasks are often complex like varieties of accents, the background noise. The paper 'Adaptable Multi-Domain Language Model for Transformer ASR' applies adaptor to the transformer ASR to promote the performance of the multi-domain problem [11]. Using three different LMs during the experiment and the input representation reach up to 353M utterances and test cases divided into three Domains. The results show the through iterations, the LMs tend to get similar accuracy, WER

together with the function to adapt to some specific nouns. Based on the huge amounts of data, the multi-domain structure is accessible which definitely reveals its potential.

As a crucial part of Spoken Dialog Systems, Spoken Language Understanding Systems (SLU) parse spoken utterances into corresponding semantic structures [12]. The ASR transcriptions, the first step of SLU are often noisy and erroneous and strongly affect the downstream processing. But applying N-Best ASR Transformer to word lattice get a better performance especially under the circumstances that the fewer input representation is given. Performing a hyper-parameter tuning on the validation set, it ranks the first in the Accuracy. Though in the condition of extremely low data regime, Transformer ASR achieves good performances.

3.2. Summary

There are several advantages of transformer-based ASR solutions.

Firstly, it possesses parallel processing capability. The parallelization ability of the Transformer model significantly reduces training time, especially crucial for ASR systems that need to process large amounts of data.

Secondly, it could be applied to capture long-distance dependencies. Thanks to the self-attention mechanism, the Transformer can effectively capture dependencies in long sequences, enhancing the accuracy of speech recognition.

Thirdly, it has superior flexibility and scalability. The Transformer is easily adaptable to different tasks and domains, and its model size and structure can be adjusted for specific applications.

However, there are several limitations.

Firstly, it has high resource consumption. Despite the significant parallel processing capabilities of the Transformer during training, it demands high computational resources, particularly when handling very large datasets.

Secondly, its training depends on large amounts of data. To achieve optimal performance, the Transformer generally requires extensive annotated data, which may pose a challenge in some resource-limited languages or specialized domains.

The successful application of the Transformer model in the ASR domain has opened new possibilities for processing natural language and speech signals. However, to fully unlock its potential, issues regarding resource consumption and dependency on large datasets need to be addressed. Future research may focus on optimizing the structure and training methods of the Transformer model to reduce resource consumption and enhance efficiency.

Moreover, by combining different types of neural network models, such as integrating the Transformer with LSTM, the advantages of both can be utilized, providing a more powerful and adaptable solution for ASR. With ongoing technological progress and continuous innovation in the deep learning field, Transformer-based ASR systems are expected to find broader applications in the future, paving new paths for human-machine interaction.

In general, the Transformer training is more stable compared to the LSTM, although it also seems to overfit together with more problem of generalization, while LSTM model requires less time to achieve the same level consequence. The subsequent modification will help the Transformer to gain more promotion [13].

4. Conclusion

Deep learning, automatic feature detection, big data processing show great applicability to the ASR. As the result of them, the LSTM model get impressive achievement. The statement is further reinforced by two significant studies on speech and feature enhancement by Weninger, Felix and others. In the meanwhile, Dong, L., and others first demonstrated Transformer-based's application in

ASR, which was subsequently adopted in end-to-end ASR models to improve the capture of contextual information. Getting hold of the basic characteristics of the two model, Karthik Ganesan and others give the probability of the Transformer-based to get less input but train the model efficiently. To sum all, LSTM and Transformer based ASRs gain great development in the region and boarden the boundary of the possible speech recognition. Their ongoing evolution, marked by research efforts aimed at overcoming their respective limitations, is expanding the horizons of what's achievable in speech recognition. The convergence of these models with other AI domains, such as attention mechanisms and multimodal integration, signifies a future where ASR systems are more accurate, efficient, and adaptable to various real-world applications, further bridging the gap between human and machine communication. In essence, the journey of LSTM and Transformer models in ASR reflects a broader trend in AI research: a move towards systems that are not just powerful in computational terms but are also nuanced and sophisticated enough to tackle the complex subtleties of human language.

References

[1] Rabiner, Lawrence, and Biinghwang Juang. An introduction to hidden Markov models. IEEE ASSP magazine, 1986, 3(1): 4-16.

[2] Bahl, Lalit R., Frederick Jelinek, and Robert L. Mercer. A maximum likelihood approach to continuous speech recognition. IEEE transactions on pattern analysis and machine intelligence, 1983, 2: 179-190.

[3] Hinton, Geoffrey, Li Deng, Dong Yu, George E. Dahl, Abdel-rahman Mohamed, Navdeep Jaitly, Andrew Senior et al. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. IEEE Signal processing magazine, 2012, 29(6): 82-97.

[4] Van Houdt, Greg, Carlos Mosquera, and Gonzalo Nápoles. A review on the long short-term memory model. Artificial Intelligence Review, 2020, 53(8): 5929-5955.

[5] Zeng, Taiyao. Deep Learning in Automatic Speech Recognition (ASR): A Review. In 2022 7th International Conference on Modern Management and Education Technology, 2022: 173-179.

[6] Weninger, Felix, Hakan Erdogan, Shinji Watanabe, Emmanuel Vincent, Jonathan Le Roux, John R. Hershey, and Björn Schuller. Speech enhancement with LSTM recurrent neural networks and its application to noise-robust ASR. In Latent Variable Analysis and Signal Separation: 12th International Conference, 2015, 12: 91-99.

[7] Narayanan, Arun, and DeLiang Wang. The role of binary mask patterns in automatic speech recognition in background noise. The Journal of the Acoustical Society of America, 2013, 133(5): 3083-3093.

[8] Weninger, Felix, Jürgen Geiger, Martin Wöllmer, Björn Schuller, and Gerhard Rigoll. "Feature enhancement by deep LSTM networks for ASR in reverberant multisource environments." Computer Speech & Language 28, no. 4 (2014): 888-902.

[9] Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. Advances in neural information processing systems, 2017, 30: 1-11.

[10] Dong, Linhao, Shuang Xu, and Bo Xu. Speech-transformer: a no-recurrence sequence-to-sequence model for speech recognition. In 2018 IEEE international conference on acoustics, speech and signal processing, 2018: 5884-5888.

[11] Lee, Taewoo, Min-Joong Lee, Tae Gyoon Kang, Seokyeoung Jung, Minseok Kwon, Yeona Hong, Jungin Lee et al. Adaptable multi-domain language model for transformer asr. In 2021 IEEE International Conference on Acoustics, Speech and Signal Processing, 2021: 7358-7362.

[12] Ganesan, Karthik, Pakhi Bamdev, Amresh Venugopal, and Abhinav Tushar. N-best ASR transformer: Enhancing SLU performance using multiple ASR hypotheses. ArXiv Preprint, 2021: 2106.06519.

[13] Zeyer, Albert, Parnia Bahar, Kazuki Irie, Ralf Schlüter, and Hermann Ney. A comparison of transformer and lstm encoder decoder models for asr. In 2019 IEEE Automatic Speech Recognition and Understanding Workshop, 2019: 8-15.