Introduction

Introducing a groundbreaking approach: Adapting Large Language Models with Speech for Fully Formatted End-to-End Speech Recognition. This innovative system integrates powerful language models with speech recognition technology, offering unparalleled accuracy and efficiency. By harnessing the capabilities of large language models, it achieves seamless transcription of spoken language into fully formatted text, revolutionizing the field of speech recognition. Say goodbye to cumbersome post-processing—this solution delivers end-to-end transcription with remarkable precision, paving the way for streamlined workflows and enhanced user experiences. Embrace the future of speech recognition with this cutting-edge adaptation of large language models.

Motivation

Conventional Automatic Speech Recognition (ASR) systems typically generate raw text that lacks proper formatting, such as punctuation and capitalization, necessitating additional post-processing steps. This post-processing often fails to effectively utilize acoustic cues present in the speech, such as intonation and pauses, which can indicate sentence boundaries and other formatting elements. The goal is to harness the power of Large Language Models (LLMs) to address these shortcomings. By integrating LLMs with ASR, the system aims to directly produce fully formatted transcriptions, eliminating the need for separate post-processing. This approach leverages the advanced contextual understanding of LLMs to interpret and apply formatting based on the nuances of spoken language, resulting in more accurate and user-friendly transcriptions.

Dataset

The SPGISpeech Dataset is a comprehensive collection designed for advanced speech recognition tasks, particularly in the financial sector. It includes 5,000 hours of financial audio recordings, meticulously transcribed to provide high-quality, fully formatted text, ensuring accuracy and detail. This dataset features a variety of speaker accents, both native (L1) and non-native (L2), enhancing its robustness and applicability in real-world scenarios. However, a key limitation is its domain-specific nature, focusing exclusively on financial contexts. This specialization means the dataset may not generalize well to other fields or casual conversational speech, potentially limiting its broader applicability in general speech recognition tasks.

Related Work

The adaptation of large language models for end-to-end (E2E) speech recognition (ASR) leverages text-only data to enhance transcription accuracy, making use of vast text corpora to improve language understanding without needing additional speech data. Techniques such as language model (LM) fusion—including shallow fusion, cold fusion, and deep fusion—integrate

pre-trained language models with ASR systems at various stages to optimize performance. A significant challenge lies in aligning speech and text representations due to their different modalities, requiring sophisticated algorithms to bridge the gap effectively. Historically, formatting of transcriptions was handled post-processing, adding complexity and reducing efficiency. This novel approach aims to integrate formatting directly into the E2E ASR process, streamlining operations and providing cleaner, more accurate outputs.

Proposed Approach

Adapting pretrained Large Language Models (LLMs) for fully formatted end-to-end (E2E) Automatic Speech Recognition (ASR) involves leveraging their advanced linguistic capabilities to directly generate formatted text from speech input. This is achieved using a composable model architecture, where a speech encoder is seamlessly integrated with a pretrained LLM. The architecture can take the form of either encoder-decoder or decoder-only structures, providing flexibility in model design. The primary aim is to make minimal architectural changes to facilitate efficient integration, ensuring that the robust language understanding of LLMs enhances the ASR system without extensive modifications. This approach streamlines the transcription process, producing accurate and well-formatted text outputs directly from spoken language.

Encoder-Decoder Based LLM

This approach integrates the speech encoder output with a text decoder, enabling seamless translation of spoken language into written text. It utilizes Connectionist Temporal Classification (CTC) for down-sampling the speech representation, enhancing the efficiency and accuracy of the model. To optimize performance, three types of loss functions are employed: CTC for aligning the speech and text sequences, Cross-Entropy (CE) for training the decoder on labeled data, and Masked Language Modeling (MLM) for leveraging large language model pretraining. This combination of techniques ensures robust handling of diverse speech inputs, precise transcription, and context-aware text generation, resulting in a highly effective end-to-end speech recognition system.

Figure

The image illustrates the composition of a speech encoder and an encoder-decoder-based large language model (LLM) for end-to-end speech recognition. The model architecture includes three main components: a speech encoder, a text encoder, and a text decoder. The speech encoder processes the input speech signal, converting it into a sequence of representations. The text encoder handles the text input, converting it into representations that the model can understand. The text decoder then combines the outputs of both the speech and text encoders to generate the final text output. The model employs both Connectionist Temporal Classification (CTC) and Cross Entropy/Masked Language Modeling (CE/MLM) for training. CTC aligns the speech encoder's output with the target text, while CE/MLM is used for the text decoder to

improve the language modeling capabilities. This approach aims to produce fully formatted text from speech inputs efficiently.

Decoder-only based LLM

This innovative approach leverages GPT-series models, such as GPT-2, to enhance speech recognition capabilities. By using a speech encoder, the outputs are transformed into prompts that the large language model (LLM) can process, facilitating accurate transcription. To tackle the common issue of length mismatch between spoken input and textual output, a CTC (Connectionist Temporal Classification)-based down-sampling method is employed, ensuring more precise alignment. Remarkably, this method requires only minimal modifications to the existing pretrained LLM architecture, preserving the integrity and performance of the original model while extending its functionality to handle speech inputs effectively. This seamless integration streamlines the end-to-end speech recognition process, providing fully formatted text outputs with enhanced efficiency and accuracy.

Figure

The image depicts the architecture of a decoder-only based large language model (LLM) for end-to-end speech recognition. The model comprises two main components: a speech encoder and a language model (LM). The speech encoder processes the input speech signal, converting it into a sequence of features. These features are then down-sampled before being fed into the language model. The language model takes these down-sampled features and produces the final text output. The model employs Connectionist Temporal Classification (CTC) and Cross Entropy (CE) for training. CTC aligns the speech encoder's output with the target text, ensuring accurate transcription, while CE helps in training the language model for better language understanding. This approach simplifies the architecture by using a decoder-only model, focusing on efficient and accurate end-to-end speech-to-text conversion.

Experimental Setup

This pioneering system leverages a vast training dataset comprising 75,000 hours of transcribed speech across diverse domains, ensuring comprehensive coverage of linguistic nuances and speech patterns. Crucially, it preserves raw text without preprocessing, enabling the model to learn directly from natural speech data, thereby enhancing its adaptability and robustness. For evaluation, the system utilizes the ESB benchmark's dialogistic dataset, meticulously modified to maintain orthographic consistency, ensuring fair and accurate assessments of its performance. By harnessing this extensive and meticulously curated data, the system achieves unparalleled accuracy and effectiveness in fully formatted end-to-end speech recognition, marking a significant advancement in the field.

Baseline Model

This pioneering approach combines an attention-based encoder-decoder (AED) model, optimizing both encoding and decoding processes for robust speech recognition. The encoder comprises 24 conformer layers, each equipped with 8-head attention mechanisms, ensuring comprehensive feature extraction from input speech data. Meanwhile, the decoder incorporates 6 transformer layers, facilitating efficient decoding of encoded information. To enhance accuracy, the system leverages a combination of Connectionist Temporal Classification (CTC) and attention Cross-Entropy (CE) losses, enabling effective alignment of predicted and ground-truth transcripts. This synergy between advanced architectural components and loss functions results in unparalleled performance, enabling fully formatted end-to-end speech recognition with exceptional precision and efficiency.

Adaptation Techniques

This innovative approach involves two distinct adaptation strategies. First, through Encoder-Decoder Adaptation, the pretrained Z-Code++ model is utilized, employing both Connectionist Temporal Classification (CTC) and Cross-Entropy (CE) losses on paired data for effective training. Additionally, Masked Language Model (MLM) loss is applied solely on text data, enhancing model understanding. On the other hand, Decoder-Only Adaptation leverages GPT-2 models integrated with LoRA adapters. These models are initially pretrained using CTC loss, followed by comprehensive end-to-end training. These strategies enable efficient adaptation and robust performance, bridging the gap between large language models and speech recognition systems for enhanced accuracy and usability.

Evaluation Metrics

The evaluation metrics for our proposed system revolve around Token Error Rate (TER), which gauges the disparity between predicted transcriptions and the ground-truth text. It encompasses both punctuation and case sensitivity, ensuring a comprehensive assessment. Furthermore, our evaluation includes a comparative analysis with existing public Whisper models, providing valuable insights into the performance of our adapted large language model in the context of speech recognition. TER serves as a robust indicator of accuracy, shedding light on the efficacy of our system in transcribing speech into fully formatted text. By benchmarking against established models, we can validate the advancements and potential superiority of our approach, thereby affirming its viability for real-world applications in speech recognition.

Experimental Results

This pioneering system's performance has been rigorously assessed across diverse datasets including CommonVoice, GigaSpeech, LibriSpeech, SPGISpeech, TED-LIUM, and VoxPopuli. The findings showcase significant enhancements in Translation Error Rate (TER) across

multiple domains following the adaptation of Large Language Models (LLMs). These results underscore the adaptability and effectiveness of LLMs in enhancing speech recognition accuracy. By leveraging the power of LLMs, the system achieves remarkable improvements in transcription quality, transcending domain-specific challenges. This comprehensive evaluation highlights the versatility and robustness of the proposed approach, promising advancements in end-to-end speech recognition technology.

Key Findings

Key findings indicate that adapting pre-trained Large Language Models (LLMs) significantly improves the performance of Fully Formatted End-to-End Automatic Speech Recognition (E2E ASR). Both encoder-decoder and decoder-only structures prove to be effective in this adaptation process. Additionally, employing CTC-based down-sampling techniques proves successful in mitigating issues related to length mismatch, enhancing the overall accuracy and efficiency of the system. These findings underscore the versatility and effectiveness of adapting LLMs for fully formatted E2E ASR, offering valuable insights for advancing speech recognition technology.

Conclusion

This study marks a groundbreaking advancement in the realm of automatic speech recognition (ASR) by successfully integrating pretrained large language models (LLMs) to enhance fully formatted transcription accuracy. By doing so, it unveils the potential to significantly enhance the readability and usability of ASR transcriptions, promising clearer and more coherent output. Moving forward, future research avenues could focus on exploring the applicability of these techniques across diverse domains and developing generalizable ASR models capable of handling varied speech contexts. This pioneering work sets the stage for a new era of ASR technology, where advancements in LLM adaptation hold the key to more accurate, accessible, and user-friendly speech recognition systems.