NLP/ASR

Unit 5: Advanced Topics in ASR/NLP

5.3.1

Advanced ASR Models

Self-supervised, Supervised, and Streaming Models

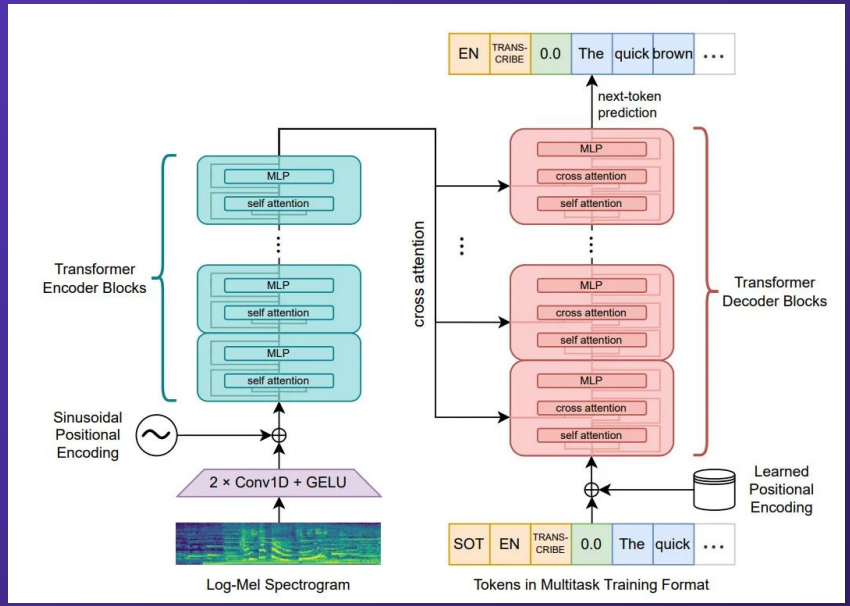
TIL-AI
TODAY I LEARNED AI

Supervised models

- Supervised ASR models are trained on large datasets of audio paired with corresponding text transcripts
- This labeled data teaches the model to associate speech patterns with specific words and phrases
- High-quality, diverse datasets are crucial for supervised ASR model performance

OpenAI Whisper

- OpenAI's Whisper is a robust, open-source ASR model
- Trained on a massive 680,000 hours of multilingual data, exceeding the scale of most ASR models
- Whisper demonstrates exceptional accuracy, handling accents, noise, and technical language
- It also excels in multilingual transcription and translation from various languages into English



Whisper

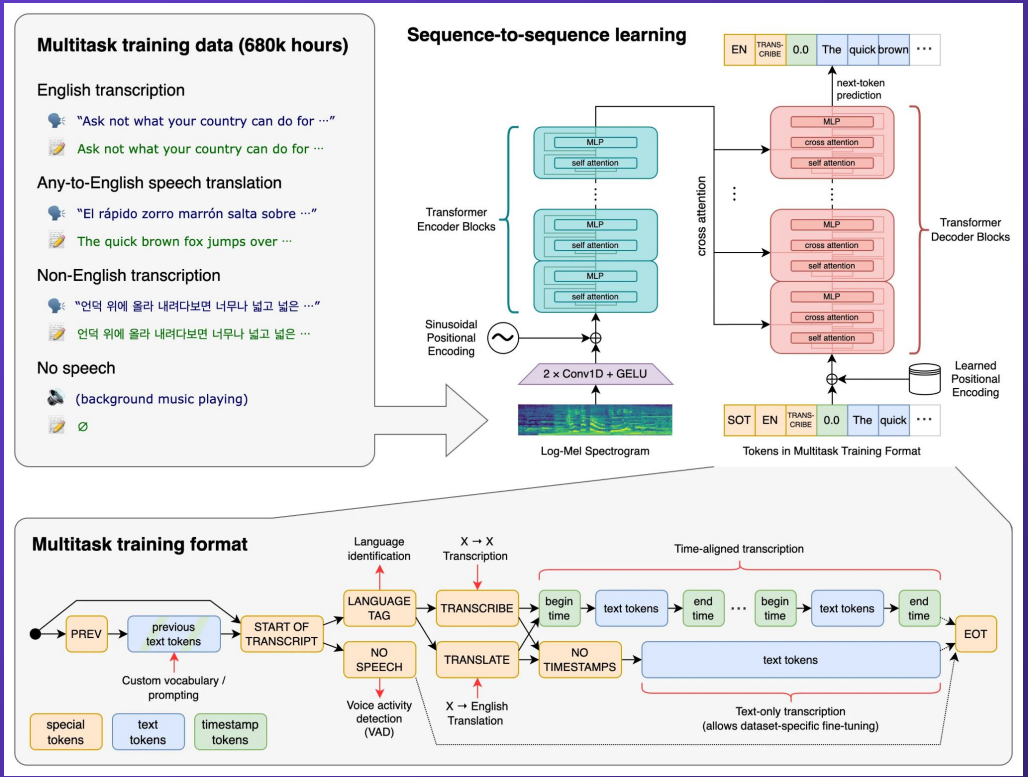
TIL-AI
TODAY I LEARNED-AI

Whisper Architecture

- Core Concept: Encoder-Decoder Transformer

- Key Components:
  - Audio Preprocessing
  - Encoder
  - Decoder
  - Special Tokens and Multitasking

Whisper Architecture

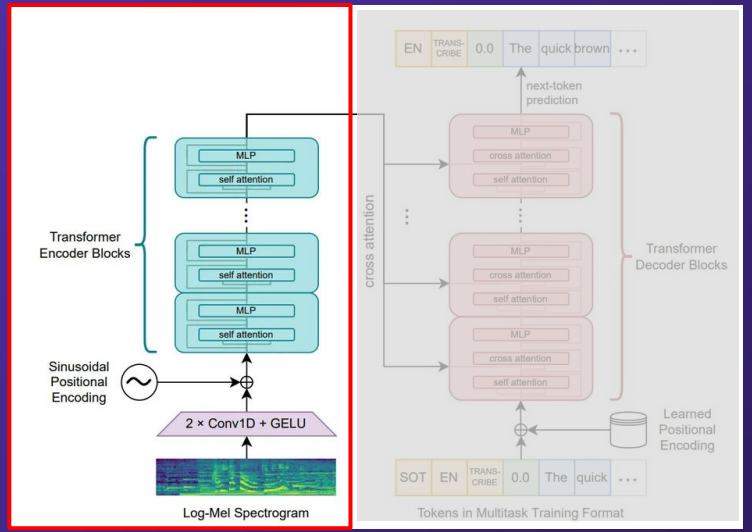


Audio Preprocessing

- Divides the raw input audio into 30-second chunks for efficient processing
- Each chunk is then converted into a log-Mel spectrogram
- Whisper can translate non-English audio directly into English text, having learned bilingual mappings during training on multilingual datasets

Encoder

- Encoder:
  - The prepared log-Mel spectrogram is fed into the encoder portion of the transformer
  - The encoder's function is to analyze the spectrogram and extract significant features that convey the speech information
  - Multiple layers with self-attention mechanisms
  - Self-attention allows the model to identify how different parts of the spectrogram (representing various frequencies and timings) relate to each other, building a comprehensive understanding of the audio content

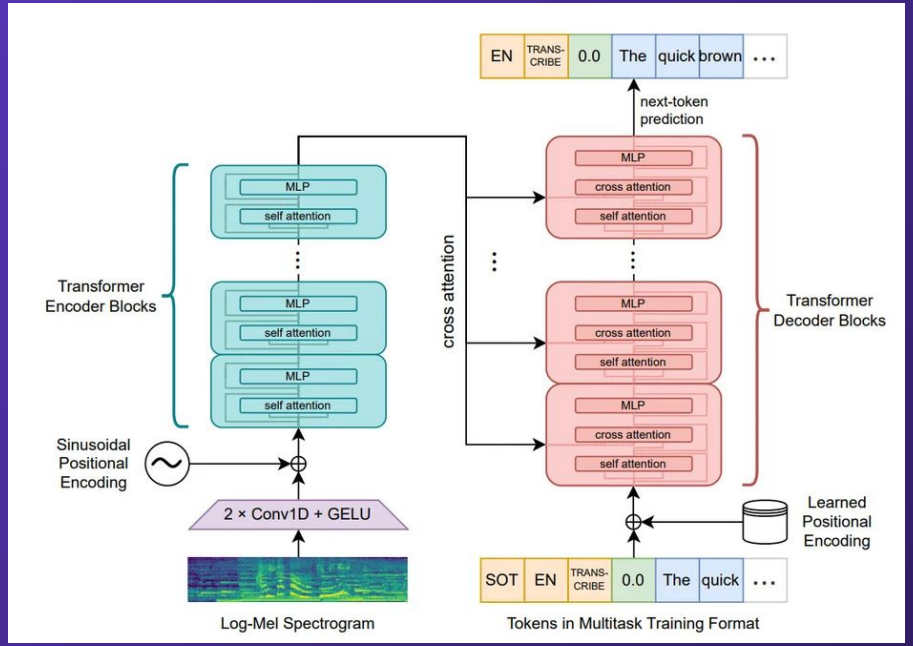


Encoder

TIL-AI
TODAY I LEARNED-AI

Decoder

- Processed information from the encoder is passed on to the decoder
- Interprets the encoded representation and generate the corresponding text transcript
- Decoder utilizes layers with attention mechanism focusing not only on the encoder's output but also on previously generated parts of the transcript itself. This context-aware decoding ensures the generated text aligns coherently word by word



Special Tokens and Multitasking

- Whisper's decoder incorporates special tokens alongside regular text tokens

- Special tokens empower the model to perform additional tasks concurrently such as:
  - **Language identification**: Tokens can signal the language being spoken, enabling multilingual transcription
  - **Phrase-Level timestamps**: Tokens can indicate timestamps within the transcript, marking the start and end of specific phrases within the audio
  - **To-English speech translation**: Special tokens can guide the model to translate the transcribed speech into English

Self-supervised models

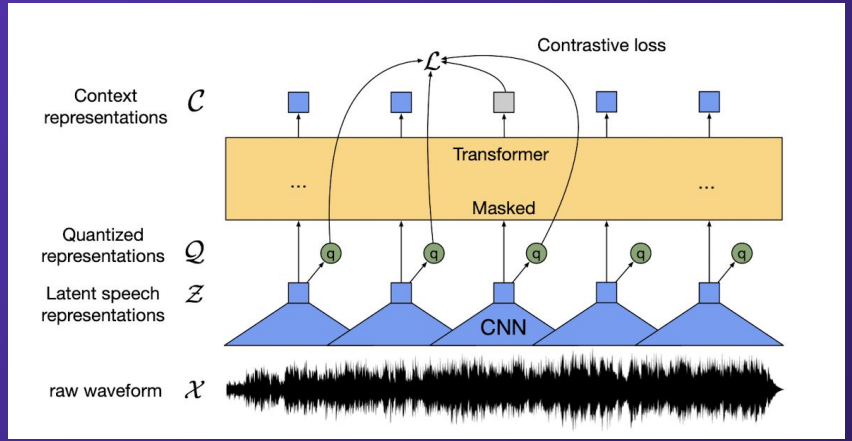
- Traditional ASR systems depend on extensive datasets of transcribed speech, limiting their use for less common languages or specific domains
- Self-supervised ASR methods aim to train models directly from raw audio without labeled data
- This approach opens up numerous possibilities in building speech recognition for low-resource languages or specialized domains

Self-supervised models

- The way self-supervised models work can be broken down into two stages:
    - Pre-training: The model learns by itself to understand and process raw audio data without any labels or transcriptions. This is similar to immersing yourself in a language, picking up on the sounds and how they connect. The model is not yet trying to understand words or sentences.
    - Fine-tuning: After the model has learned a good representation of audio from the pre-training, it is fine-tuned with transcribed speech. The goal here is to teach the model to accurately map the audio it hears to the correct words written in the text. The model learns to understand and transcribe spoken language into text.

Wav2Vec 2.0

- Wav2Vec is a breakthrough framework for self-supervised pre-training of speech representation models
- It uses a combination of convolutional and transformer-based neural networks
- Key idea - Contrastive learning forces the model to distinguish true audio segments from distorted versions
- Wav2Vec 2.0 extends the original Wav2Vec by moving towards full speech-to-text functionality

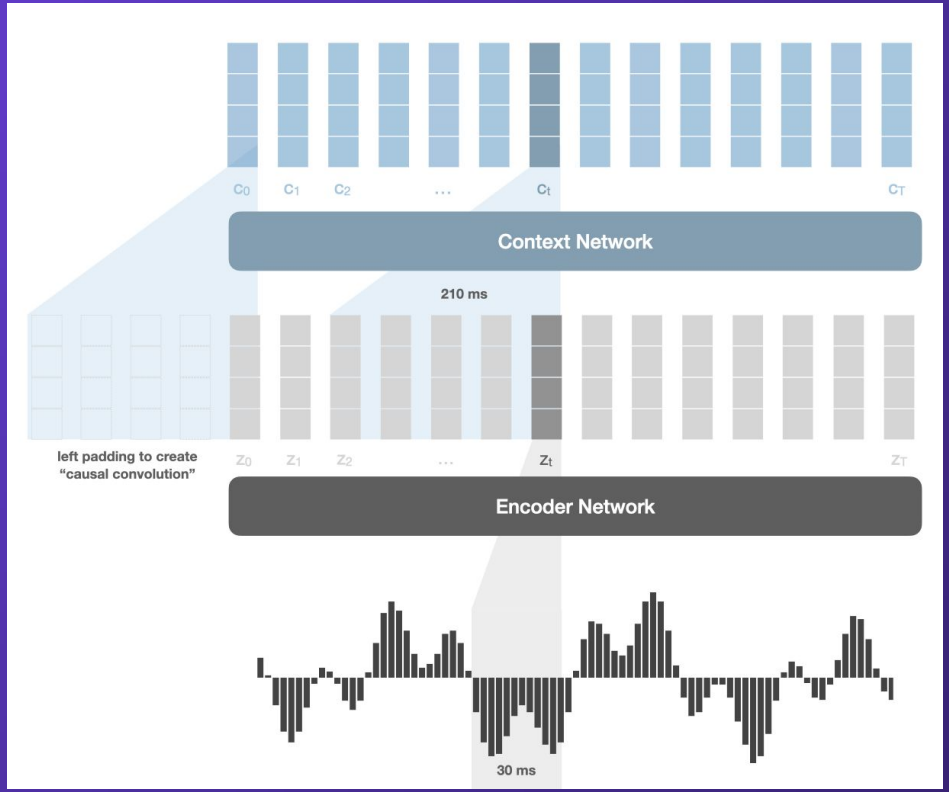


Wave2Vec 2.0

Wav2Vec 2.0 Architecture

- Key Components:
  - Feature Encoder
  - Context Network
  - Quantization Module
  - Contrastive Predictive Coding

Dual Network Backbone



TIL-AI
TODAY I LEARNED AI

Dual Network Backbone

- Encoder Network:
  - (CNN) responsible for reducing raw audio dimensionality
  - Processes small chunks of audio (typically 30 milliseconds) and outputs a lower-dimensional feature vector (around 512 dimensions) for each chunk, capturing essential audio characteristics

- Context Network (Transformer Layer):
  - Another CNN that takes the encoder's output features as input
  - Capture long-range dependencies within the audio.
  - Takes larger context window (often 210 milliseconds) and outputs another 512-dimensional feature vector for each chunk, summarizing the information within that extended context
  - Uses self-attention mechanisms to weigh the importance of different parts of the audio sequence when predicting a particular element, making it powerful for modeling sequences.

Quantization

- A novel scheme to learn discrete speech units
- Maps the dense representations from the CNN into a set of discrete latent variables
- Utilizes codebooks, which are collections of possible codewords. The model effectively learns to represent the latent features (internal representation of the audio) using combinations of these codewords
- Helps in handling the variability in the audio signals more robustly

Contrastive Predictive Coding (CPC)

- Wav2Vec doesn't directly predict the future audio samples

- Uses contrastive loss function based on a concept called CPC
  - Focuses on maximizing the mutual information between the context representation and a future audio sample
  - Tries to learn how well the context information predicts the characteristics of a sample from a bit further ahead in the audio stream.

Real-Time ASR

- Real-time ASR converts spoken words into text as they are being spoken
- It has low latency, providing near-instantaneous transcription
- This technology is crucial for applications like live captioning, dictation, and virtual assistants

How Streaming Models Work

- Streaming models process audio input in small chunks or segments
- The model transcribes each chunk as it arrives
- Partial results are updated continuously, providing a real-time transcription experience
- Real-time processing requires efficient models and algorithms

Popular Streaming ASR Frameworks

- Whisper
- Vosk: Open-source, offline streaming ASR toolkit supporting multiple languages
- NVIDIA NeMo: Toolkit for conversational AI, including powerful streaming ASR capabilities