Qwen-Audio

Abstract

- Các mô hình trước hạn chế trong khả năng tương tác vì bị giới hạn data(chỉ tiếng nói người, …)
- Qwen-Audio: chạy trên 30 task khác nhau
- Khó khăn: dễ bị noise khi train 30 task
- Khắc phục:
  - Thiết kế một framework đào tạo đa tác vụ thông qua decoder

Introduction

- Khuyến khích knowledge share, giảm thiểu interference, specified tags
- Kết hợp speech recognition và word-level time-stamp prediction (SRWT) task => cải thiện QA và improves the performance of ASR
- Qwen-Audio-Chat
  - supervised instruction fine-tuning => linh hoạt trong việc nói hoặc viết

Methodology

- Training bằng 2 giai đoạn
  - multitask pretraining
  - supervised fine-tuning

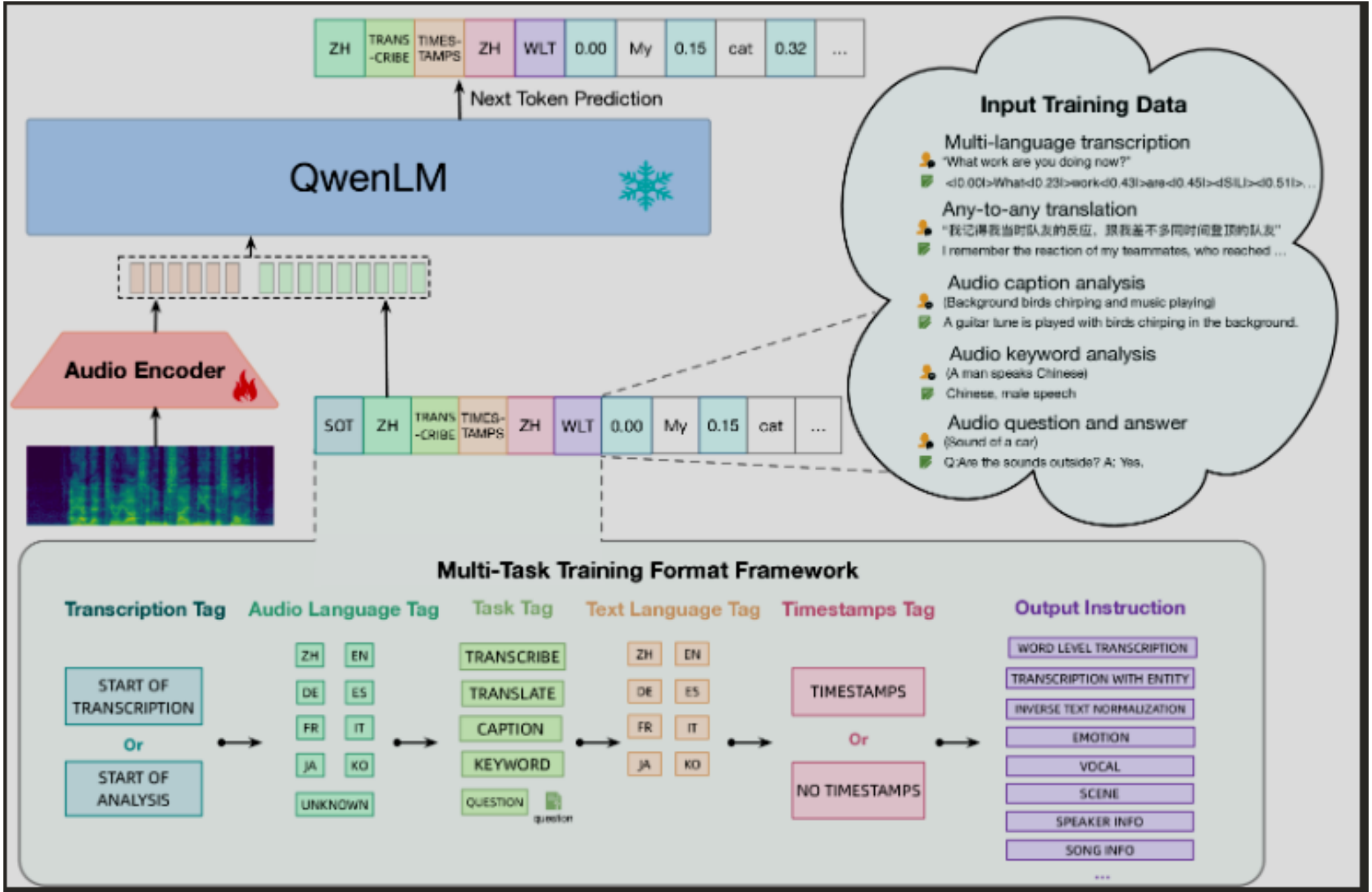

Figure 3: The overview of Qwen-Audio architecture and multitask-pretraining.

Model Architecture

- Data gồm có a: audio và x: text
- Target: maximize $\mathcal{P}_\theta(x_t \mid \boldsymbol{x}_{<t}, \text{Encoder}_\phi(\boldsymbol{a})),$
- **Audio Encoder**
  - Sử dụng Whisper-large-v2
  - SpecAugment là data augmentation
- **Large Language Model**

Multitask Pretraining

- Mục tiêu là đào tạo đồng thời để tạo thành một module thống nhất cho các task
- Lợi ích:
  - Các task gần giống nhau có thể knowledge sharing và collaborative learning
  - Các nhiệm vụ hiểu biết thấp hỗ trợ các nhiệm vụ yêu cầu hiểu biết cao

Multi-task Training Format Framework

- Transcription Tag:
  - <|startoftranscripts|>: nhận dạng đúng từ đã nói
  - <|startofanalysis|>: các nhiệm vụ khác
- Audio Language Tag: ngôn ngữ nói (8 ngôn ngữ). Nếu không có lời nói thì sử dụng: <|unknown|>
- Task Tag: <|transcribe|>, <|translate|>, <|caption|>, <|analysis|>, and <|question-answer|> tasks. Đối với QA thì thêm câu hỏi sau tag
- Text Language Tag: ngôn ngữ văn bản
- Timestamps Tag: <|timestamps|> hoặc <|notimestamps|> có cần biết thời gian hay không?
- Output Instruction: mô tả nhiệm vụ và output mong muốn

Supervised Fine-tuning

- Dành cho mô hình Qwen-Audio-Chat
- Tạo thủ công các ví dụ cho LLM (label, text, QA) -> GPT3.5 tạo thêm QA dựa trên label và text

Các token đặc biệt

- Transcription: `startoftranscript, startofanalysis`
- Task: `translate, transcribe, caption, keyword`
- Timestamp: `notimestamps, sil, timestamps`
- Language: `unknown, zh_tr(tranditional Chinese), en, zh, de, es, ko, fr, ja, it`

caption_audiocaps	Mô tả ngắn gọn audio
caption_clotho	Mô tả chi tiết
audioset_ontology	Xác định loại âm thanh
caption_plain	
itn	Inversed Text Normalized
wo_itn	không sử dụng Inversed Text Normalized
Startofentityvalue và endofentityvalue	Bắt đầu và kết thúc của entity, giúp phân loại entity (ví dụ: tên, địa điểm)
named_entity_recognition	
audio_grounding	
Startofword và endofword	Bắt đầu và kết thúc của một từ
delim	Ký tự phân cách cặp timestamp trong audio grounding
emotion_recognition	
music_description	
note_analysis	Phân tích nốt nhạc, dùng để trích xuất các thông tin về cao độ, cường độ của các nốt nhạc
pitch	Dùng trong phân tích nốt để xác định cao độ
velocity	Cường độ của một nốt
sonic	Âm sắc

instrument	Nhạc cụ
speaker_meta	thông tin về người nói, bao gồm đặc điểm giọng, giới tính, hoặc ngữ cảnh
song_meta	thông tin về bài hát, như tên, thể loại, hoặc nghệ s
Question và answer	
choice	Các lựa chọn
scene	Nhận dạng ngữ cảnh, cảnh vật trong âm thanh
event	*sound event*
vocal_classification	
speech_understanding	
scenario	*speech language understanding: scenario*
action	*speech language understanding: action*
entities	Định dạng thực thể liên quan đến ngữ cảnh hiểu ngôn ngữ
speech_edit	Định dạng chỉnh sửa nội dung lời nói, áp dụng cho các đoạn hội thoại hoặc âm thanh ngắn