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

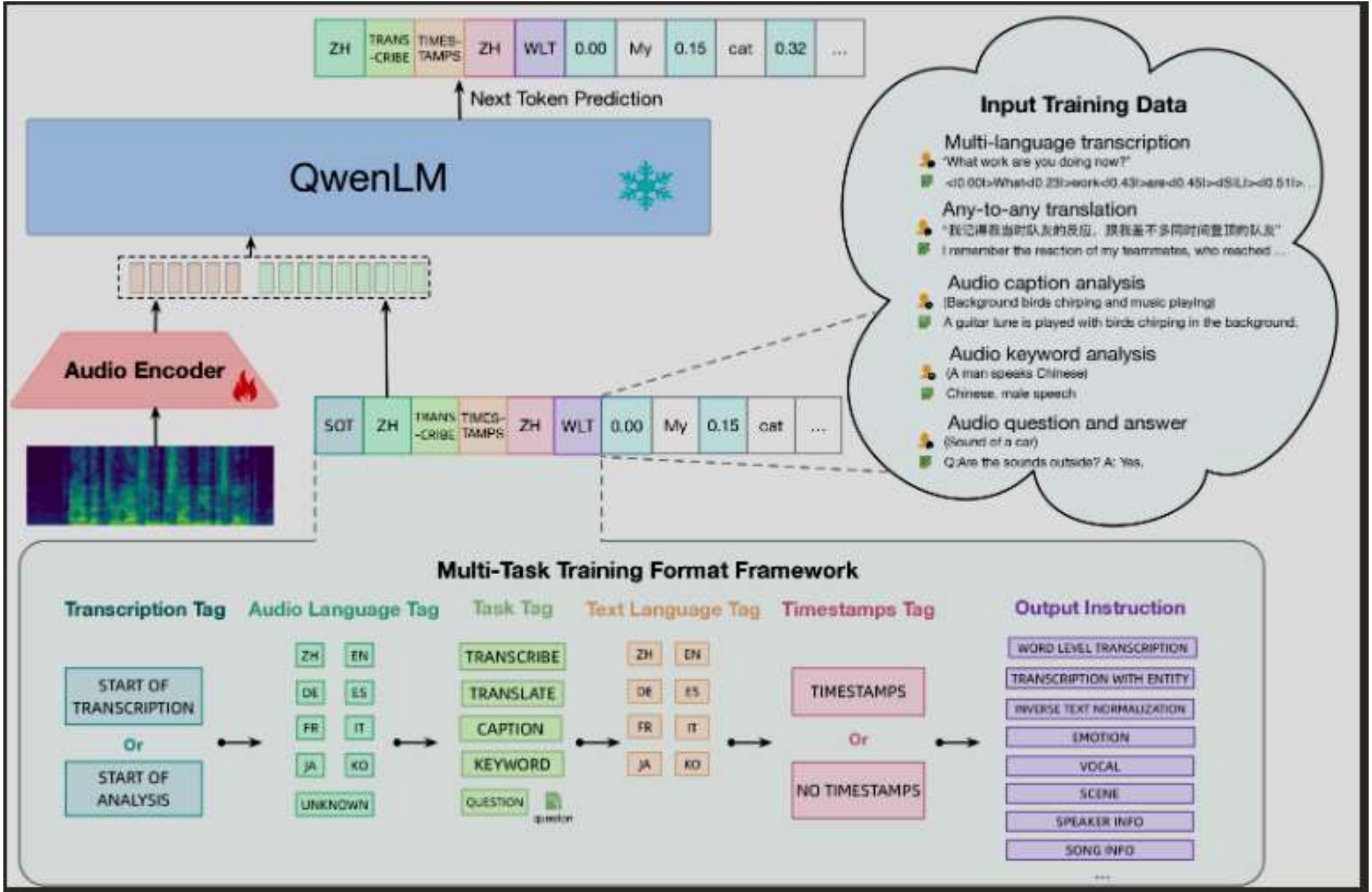

Figure 3: The overview of Qwen-Audio architecture and multitask-pretraining.

Model Architecture

- Data gồm có a: audio và x: text
- Target: maximize $\mathcal{P}_\theta(x_t \mid x_{<t}, \text{Encoder}_\phi(a)),$
- **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