Research Project in the Data Science Profile

Embedding Swiss German Audio into LLMs

Situation

The Institute for Data Science FHNW has been working on Swiss German Speech-to-Text (STT) and Text-to-Speech (TTS) models for several years. For the newest Swiss German speech recognition models, we use different versions and descendants of OpenAI's Whisper model, which has been pre-trained on vast amounts of multilingual and multitask supervised data collected from the web, covering nearly 100 different languages. We fine-tuned the model on specific Swiss German data, resulting in new state-of-the-art models; a demo can be found here: https://stt4sg.fhnw.ch/

We aim at extending the project by developing an audio embedding for Swiss German, i.e. to develop a pipeline such that the audio input can be semantically embedded as text and interpreted to generate a response downstream leveraging an LLM. This approach focusses on understanding the audio's semantic content and bypasses the step of a "perfect", literal translation into High German text for commercial applications. As a result, the latency of the chatbot's response may be substantially reduced and allow for an improved user satisfaction in practice.

Goals / Methodology / Tasks

To train a multimodal large language model for audio-to-text tasks, several steps will be needed:

- -) Audio Preprocessing: Convert raw audio into features like spectrograms or Mel-frequency cepstral coefficients for suitable input representation.
- -) Audio Encoding: Use models such as wav2vec or Whisper to encode audio features into high-dimensional embeddings that capture phonetics, speech patterns, and (wishfully) emotional tone.
- -) Multimodal Alignment: Align audio embeddings with the text-based transformer's space using cross-attention layers for focusing on relevant features or projection layers for mapping audio features.
- -) Text Generation: Use the language model to process embeddings and generate coherent textual responses.
- -) Fine-Tuning: Train the audio encoder on self-supervised tasks to understand audio patterns. Use paired audio-text datasets for task-specific learning and perform end-to-end fine-tuning by freezing the audio encoder and training specific layers.

Challenges of this project include data scarcity, noise robustness, latency, and bridging the semantic gap between temporal audio data and sequential text tokens. Data augmentation and other techniques can help to address these issues for real-world applications.

You will learn how to work with advanced AI systems for audio and language processing, apply techniques to train and fine-tune cutting-edge models and build efficient AI systems tailored for conversational applications.

Required Skills

Good programming skills, interest in deep learning and especially in Natural Language Processing (NLP), willingness to work closely in a team, knowledge of Swiss German would be an advantage.

Tasks for the Master Student

The project can cover several aspects in 2 to 3 sub-projects (IP7, IP8, IP9).

Full/Parttime: [X] Full time study
[X] Part time study

Location: Windisch

Advisor: Prof. Dr. Daniel Perruchoud