ASR + LLM for low resource languages

Kazakh language dataset is used for training purposes

Problem outline

1) **Low-resource** languages don't have enough labeled speech data for high-quality domain level Speech recognition. Ex, Kazakh language has less than 100h of open data.

2) Connectionist Temporal Classification (CTC) speech recognition models, commonly used for real-time ASR, often rely on traditional n-gram language models, which only consider 3-4 words at a time and lack full-context understanding.

Solution: Replace n-gram LM with LLM like Llama3.2

Replace LLM (LLama3.2, T5, etc) **Embedding layer** with Custom layer that takes sequence of character probabilities as an input, instead of token ids

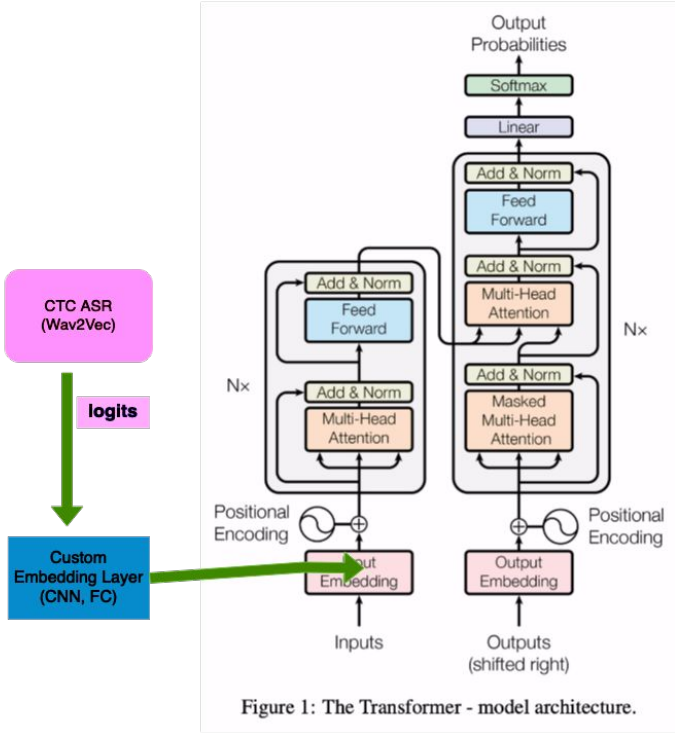

Figure 1: The Transformer - model architecture.

Benefits

1) **Low-resource** languages will be able to fine-tune LLMs on huge local unlabeled/domain-specific textual datasets, then adapt it to pre-trained ASR model (that only saw few hours of this low-resource language data).

2) **Popular languages** like English will be able to generate ASR transcriptions by taking the into account the whole context (history of conversation), not just last 3-4 words

Training results

1) **13%** word error rate (lower is better) on the validation set of English dataset (Switchboard)

2) **38%** word error rate on the validation set of **Kazakh** dataset (~4 hours). still training

Thanks!