Fine-tuned XTTS-v2 for Russian (Single Speaker)

Model description

This is a fine-tuned version of [coqui/XTTS-v2] adapted for Russian, using a single-speaker dataset.

Dataset

The dataset is based on audiobook recordings from the LibriVox project. I found it with transcriptions on the OpenSlr website. That's why there was no need to use the Whisper model to transcribe the audio, but I checked it out and tried transcribing some audio and it was successful.

I have selected single speaker audio lasting 2h in total. I have formatted the dataset in LJSpeech format as required by coqui-tts.

The audio was in a very good quality already, so there was no need to do a lot of preprocessing. I still did some basic preprocessing. I have trimmed the silences. And removed audio files with more than 182 chars, because it would cause truncation in xtts fine tuning. I have also tried removing background noise, but it seemed to worsen the quality, so I skipped this part. The audio was recorded in 16kHz but I resampled to 22050Hz.

Training

- Base model: XTTS-v2 pretrained

- Hardware: MacBook Pro M3 Max (MPS backend)

- Learning rate: 5e-6

- Epochs: ~12

- Loss: mel CE reduced from ~4.9 → ~3.9 (see plots below)

Results

The audio generated from the fine-tuned model seems to be of better quality and clearer. The base model was already producing good results, so it wasn't easy to improve a lot. Also, I only

trained on 2h of audio as this is only experimental and I didn't have a proper device for training (I borrowed a friend's laptop with M3 max).

In terms of loss stats, there was a noticeable and stable decrease.

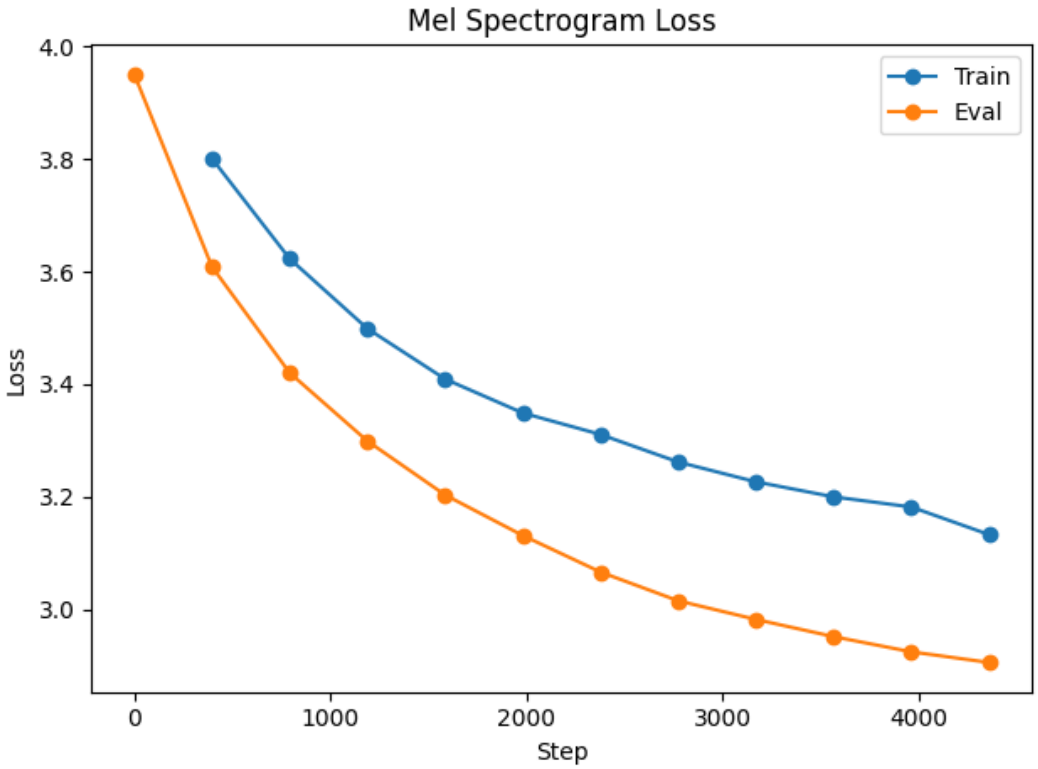

Fig 1: Mel Spectogram Loss

The fact that the validation curve closely tracks (and is slightly lower than) the training curve shows that the model generalizes well and is not overfitting. This steady reduction confirms that the fine-tuning process successfully improved the model.

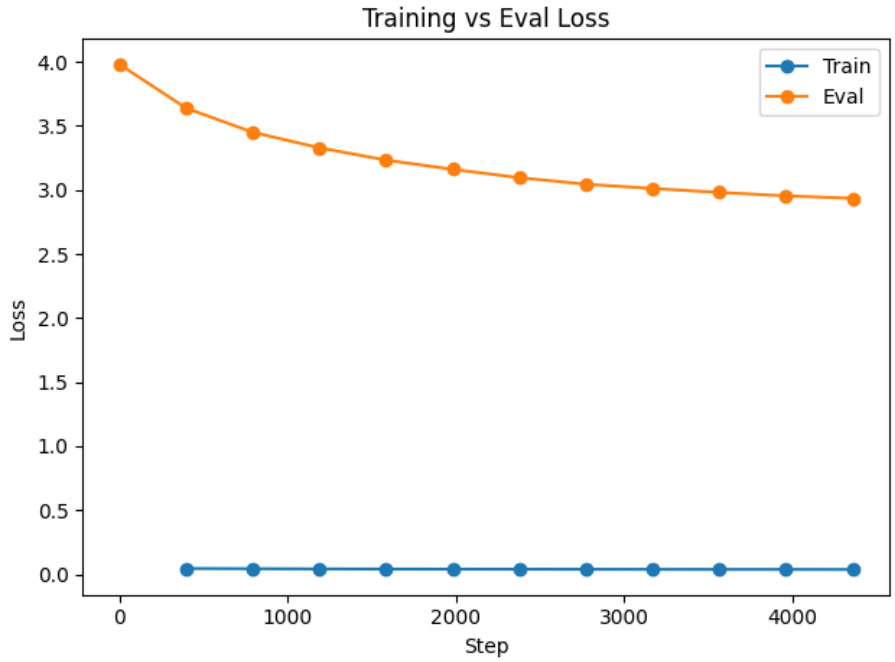

Fig 2: Training vs Evaluation loss

The training loss is very small throughout. This is expected because the model adapts to training data very well. The evaluation loss is also decreasing at a stable rate. This means that the model is generalizing well and improving in performance for unseen data.

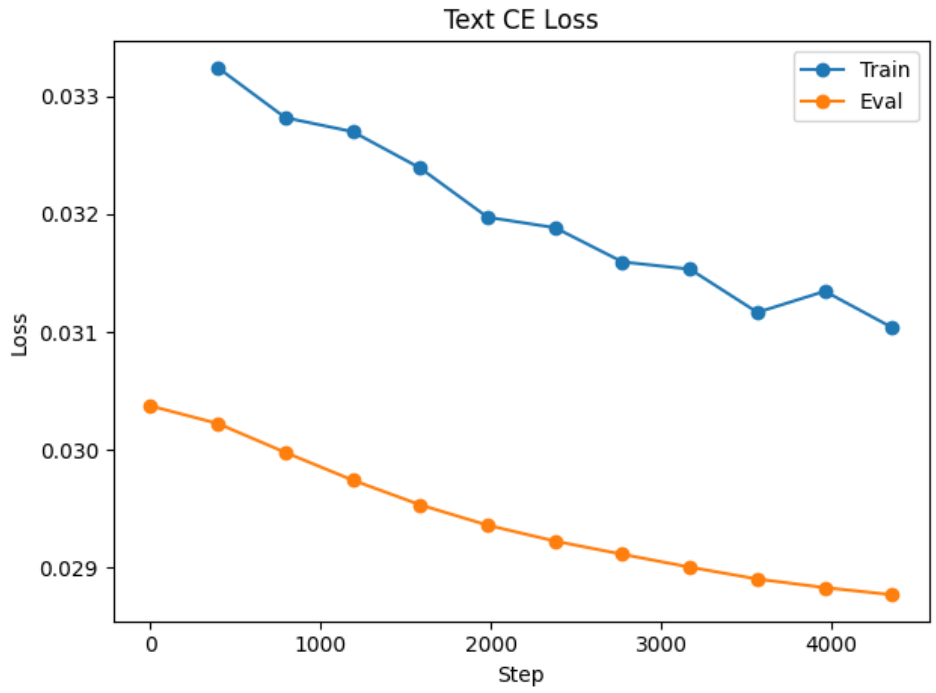

Fig 3: Text cross entropy loss

Both training and evaluation cross entropy loss is decreasing at an almost steady rate without sudden spikes. Interestingly, the eval loss is consistently lower than the training loss. This again shows that the model is generalizing well. After 4000 steps, the graph seems to be starting to flatten. This suggests that the training was almost coming to a good point to stop.

Limitations

- Single-speaker dataset

- training data is not much.