Preliminary report #2 on Semester task WS23

NVIDIA NeMo model and runtime on minicomputer

For the previous weeks, my goal was to explore the ability of Jetson Devices to run Nvidia ASR models. As mentioned in the previous report new Jetson Devices, e.g. *Jetson Xavier NX*, support NVIDIA Riva AI SDK - a solution to run a docker server on a Jetson device to request automatic speech recognition in an API manner. Riva supports NeMo models, which is useful for my experiment since the STT En Citrinet 1024 model is already ready to be used to make my measurements. Moreover, Riva supports deploying custom-trained NeMo models using nemo2riva tool.

I have improved previous code, removed redundant decimals, and altered memory measurement with python memory-profiler library.

Running the measurements code, I have got the following results on a hundred samples:

The results were surprising since *WER* was too low: `26.16`. Looking at the error examples, it turned out some of the audios produced only a have of the speech track:

Model load time (sec)	Inference time (hh:mm:ss)	Dataset samples count	Dataset length (hh:mm:ss)	RTF (%)	WER (%)	MER (%)	WIL (%)	WIP (%)	CER (%)
4	0:00:27	100	0:10:25	4.43	26.16	26.11	29.95	70.05	24.2

Listening to the error samples, I found that Riva cuts the rest of the transcription after a prolonged pause in the speech. The reason or maybe config-parameter is further to be detected. The forum-discussion suggested turning the punctuation model off, but it did not help yet.

THERE'S IRON THEY SAY IN ALL OUR BLOOD AND A GRAIN OR TWO PERHAPS IS GOOD BUT HIS HE MAKES ME HARSHLY FEEL HAS GOT A LITTLE TOO MUCH OF STEEL ANON	THERE'S IRON THEY SAY IN ALL OUR BLOOD

The other downside of Riva during the setup process was the requirement to use API-key and hence the registration in Nvidia Developer service, omitting the fact that it takes so long to initialize the river-client-server and download required models.