L46 Project Timeline

Nov 16th - Nov 23rd:

- Define project: Quantization of SpeechBrain models.
- Literature review: ASR models, quantization, SpeechBrain.

Nov 23rd - Nov 30th:

- Define ASR models: CRDNN, Wav2vec2.
- Define dataset: LibriSpeech.
- Literature review: quantization techniques, quantization frameworks.
- Define quantization techniques of interest: PTQ, QAT, Mixer precision quantization, stochastic quantization, AdaRound, Vector Quantization.

Nov 30th - Dec 7th:

- Exploration of Data, SpeechBrain and models.
- Initial experiments: develop understanding of pytorch-quantization.
- Research: integration of TensorRT, ONNX, pytorch-quantization in SpeechBrain.
- Analysis of model weights, activations, bottlenecks in architecture.

Dec 7th - Dec 14th:

- Initial experiments: PTQ of both CRDNN and Wav2Vec2.
- Implementation of missing quantized layers: LayerNorm, Embedding, GruCell.
- Export models to ONNX.

Dec 14th - Dec 21st:

- Fine tuning of CRDNN, Wav2Vec
- Implementation: QAT, Mixed precision quantization of both CRDNN and Wav2Vec2.

Dec 21s - Dec 28th:

- Experiments: PTQ, QAT, Mixed precision quantization of CRDNN and Wav2Vec2: Export to ONNX, Evaluation and analysis.

Dec 28th - Jan 5th:

- Implementation of Stochastic quantization and AdaRound for both models.
- Filling the gaps for missing results of previous experiments.

Jan 5th - Jan 12th:

- Experiments: Stochastic quantization and AdaRound for CRDNN and Wav2Vec2: Export to ONNX, Evaluation and analysis.
- Implementation of Batch Normalization Folding.

Jan 12th - Jan 18th:

- Experiments: Batch Normalization Folding for CRDNN and Wav2Vec2: Export to ONNX, Evaluation and analysis.
- Filling the gaps for missing results of previous experiments.
- Writing of the report.