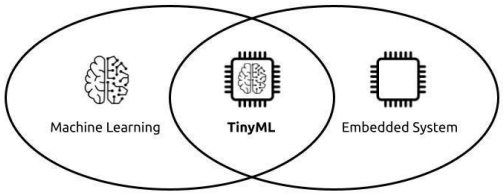
Honey I Shrunk the ML Model

Alex Melnick & Sergio Rodriguez
Boston University, College of Engineering, EC601 Project Design

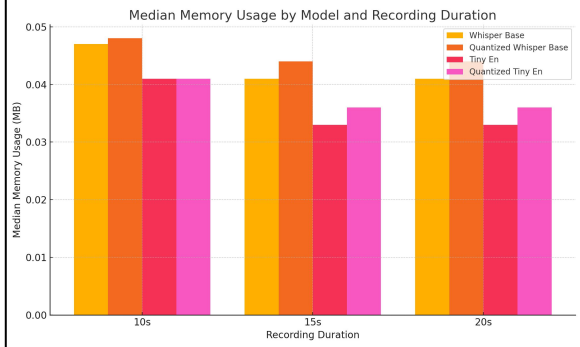
GitHub

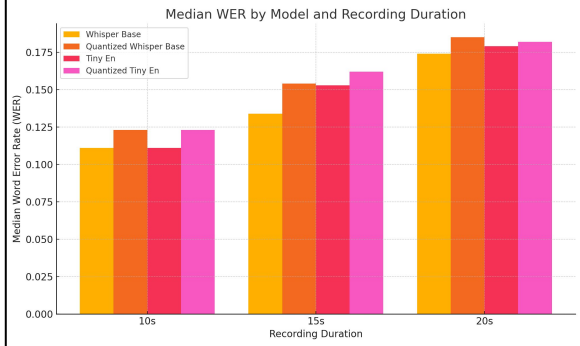
Introduction

- **Goal:** Compress Whisper Tiny EN and Whisper Base to run on microcontrollers while maintaining performance.
- **Methods:** Quantization, pruning, and TensorFlow Lite Micro conversion for edge deployment.
- **Applications:** Deaf Accessibility, IoT devices, wearables, and offline transcription.
- **Impact:** Brings state-of-the-art speech recognition to resource-constrained hardware using TinyML.



Results



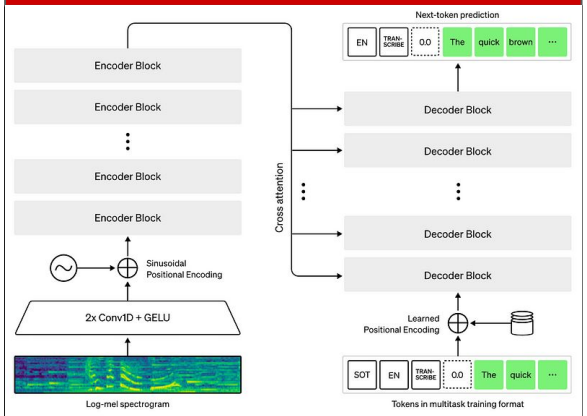


(results based on TED-LIUM)

$$\text{WER} = \frac{S + D + I}{N}$$

where...
S = number of substitutions
D = number of deletions
I = number of insertions
N = number of words in the reference

Whisper Architecture



Conclusion

We reduced the Whisper Base and Tiny En models by 48% and 23%, respectively, with only a 7% and 6% increase in memory usage and a 10% and 5% rise in WER. This optimization is a massive success for storage-constrained systems and brings us closer to enabling advanced generative AI models, such as fully on-device conversational interactions powered by LLMs and text-to-speech systems.