How Much Context Does My Attention-Based ASR System Need?

Introduction

- For the task of speech recognition, acoustic models (AMs) are usually trained on short context windows of 5-20s. This context window is generally chosen based on compute constraints.
- This work presents a study on the impact on performance of using longer context windows (of up to 1 hour) during training/evaluation.
- The aim is to examine how much context current AMs can benefit from.

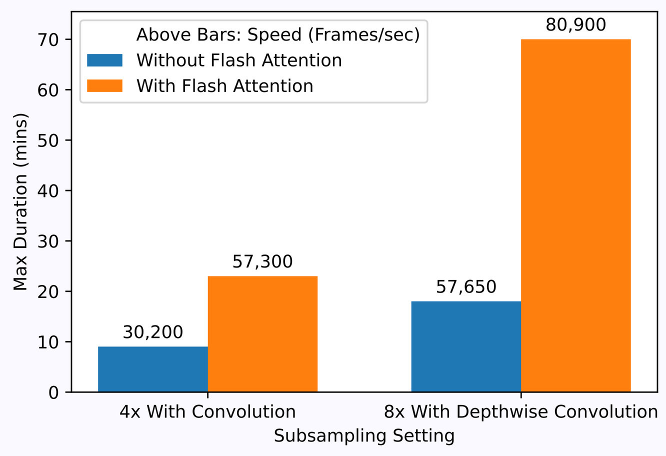
Modifications enabling training with long sequences

Architecture

- We use a CTC Conformer-Based architecture for our experiments

Flash attention is used in conjunction with 8x depthwise subsampling

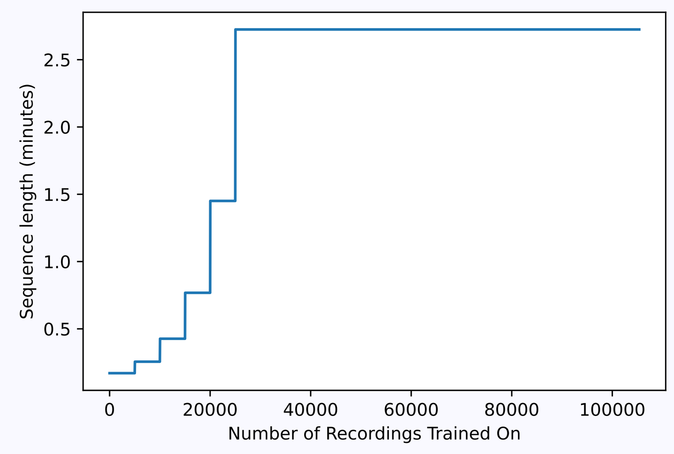
- Flash attention is a kernel for computing attention without realising the $n \times m$ attention matrix
- This combination enables **training** with contexts of up to 70 minutes on 1 A100, a magnitude larger than what is used in prior work



Sequence Length Warmup

Starting training with long sequence lengths lead to severe instability. To avoid this, the sequence length is gradually increased throughout training.

$$s_r = \min(s_0 + s_0 \cdot 2^{\lfloor r/n \rfloor}, s_m)$$

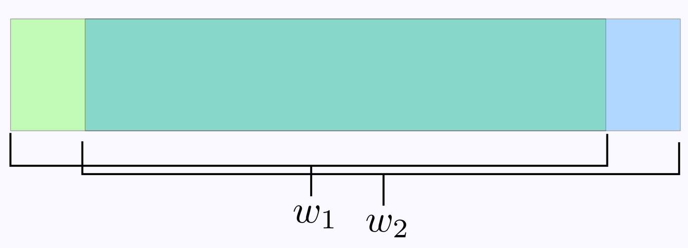


$n = 5000$, $s_0 = 5.12s$, $s_m = 1638.4s$

- s_r : Sequence length at given recording/step
- r : Recording/step index
- s_m : Maximum sequence length
- s_0 : Initial Sequence length
- n : Doubling frequency

Fairly comparing models of varying context lengths

Shorter sequence lengths result in a greater amount of context fragmentation (shown below), this causes longer context models to always perform better if evaluated naively. A moving window scheme is used to fairly compare models of different context lengths by avoiding context fragmentation.



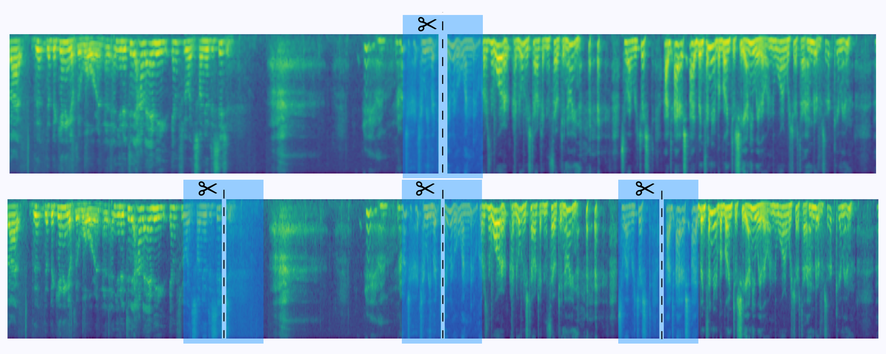


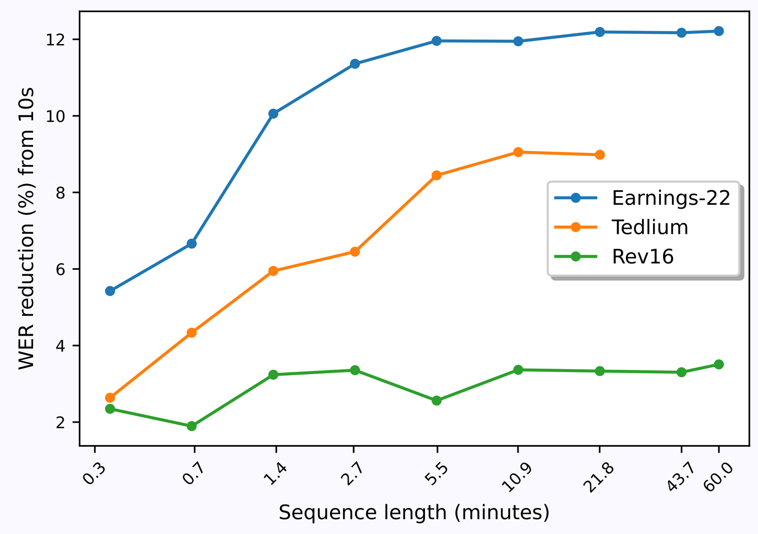
Figure 1: (Top) Sequence length of 10s (Bottom) Sequence length of 5s.

Datasets

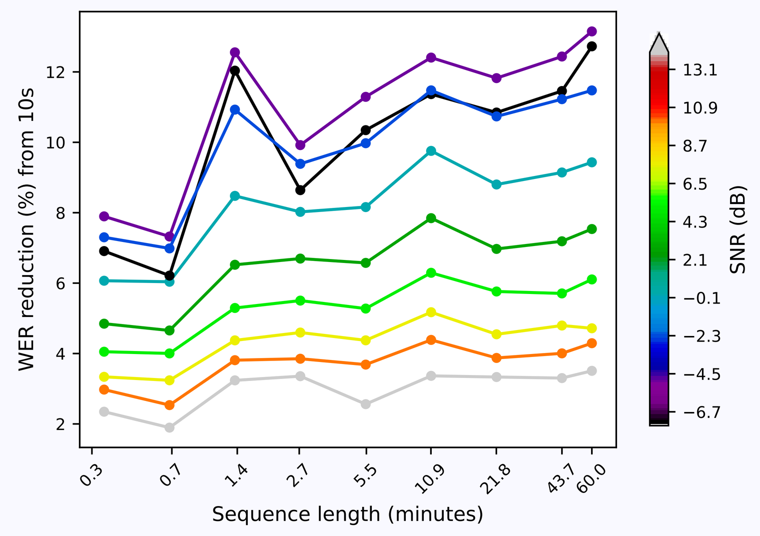
Dataset	Hours	Purpose	Max Duration (min)
Spotify Podcasts	58,000	Training	300
Tedlium	2.6	Evaluation	30
Earnings-22	119	Evaluation	123
Rev-16	16.2	Evaluation	132

How much context is useful?

The model benefits from up to 20 minutes of contexts. Earnings-22, the most challenging and out-of-domain dataset benefits the most from the context, while Rev16, out in-domain test set shows little benefit.

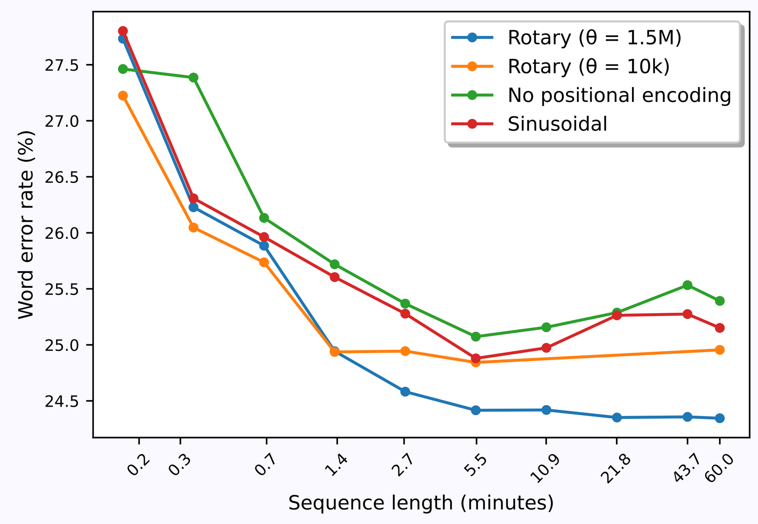


Longer context models show greater robustness to background noise. Rev-16 Dataset.



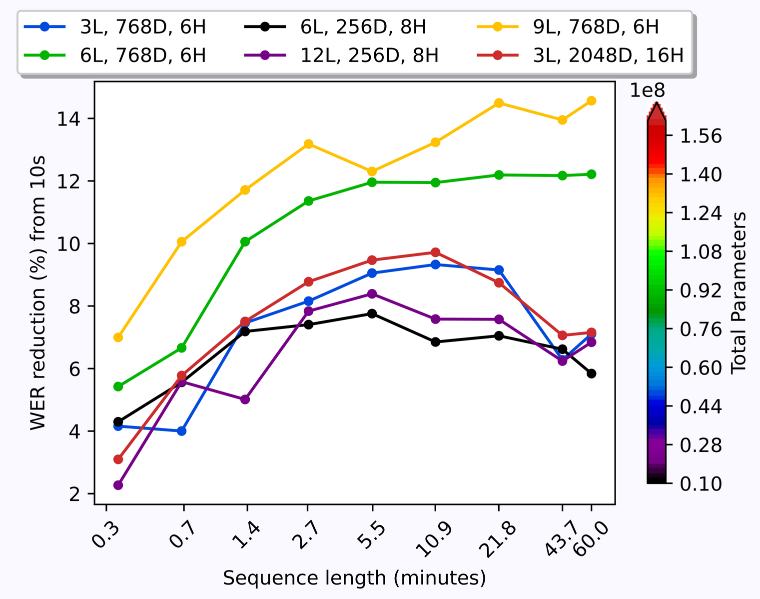
Impact of positional encoding method

Rotary Encodings lead to increasingly better performance as the context length is scaled. We find that it is crucial to increase rotary's θ parameter, which reduces the bias to nearby frames. Earnings-22 Dataset.

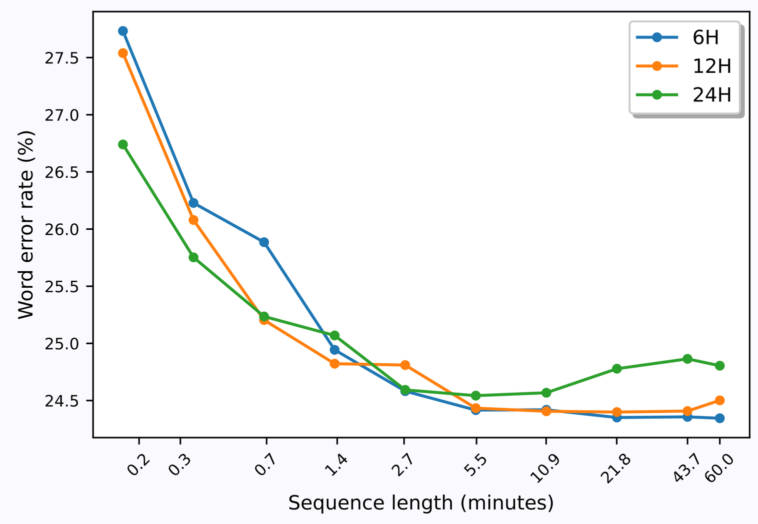


Impact of model size

Small or shallow models are less effective at robustly leveraging longer contexts. On Earnings-22 all models below 90M parameters or 6 layers showed degradation at longer contexts.



Smaller attention per-head dimensions are more effective at shorter contexts, but less effective at longer contexts. Earnings-22 Dataset.



Future Work

- **Model Interpretability** - what context features are the model benefiting from?
- **More Effective Approaches** - what approaches will enable the model to utilise a full hour of audio at test-time?

Acknowledgements

This work was supported by the Centre for Doctoral Training in Speech and Language Technologies (SLT) and their Applications funded by UK Research and Innovation [grant number EP/S023062/1].

Robert Flynn, Anton Ragni

rjflynn2@sheffield.ac.uk, a.ragni@sheffield.ac.uk

The University of Sheffield