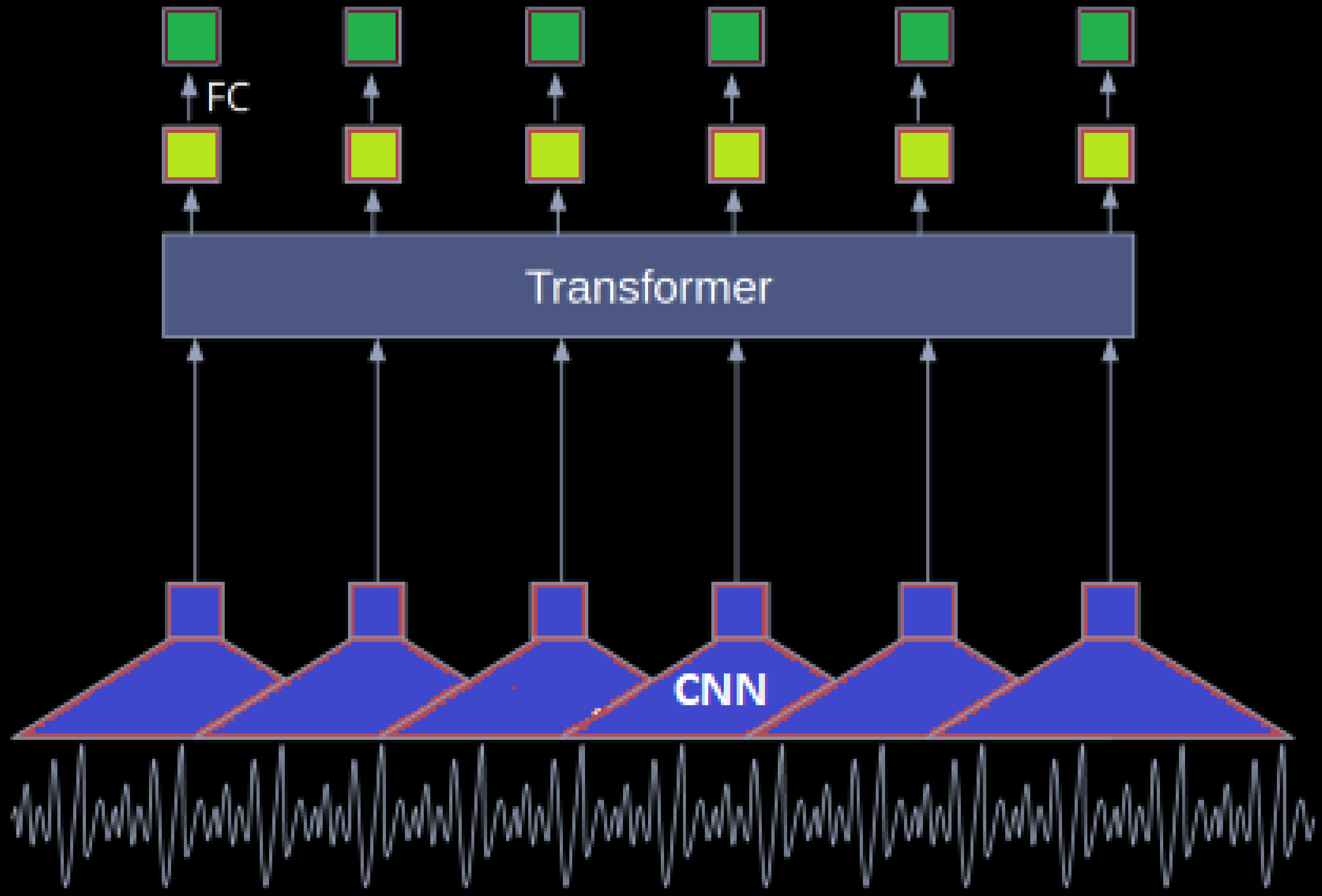
Deep Learning

Omar Farooq

Urdu Speech Recognition

Problem Statement

Urdu is the national language of 61.9 M people in the Asian Sub-continent. It lacks a freely available state of the art speech recognition system. Such a module would enable numerous natural language processing applications.

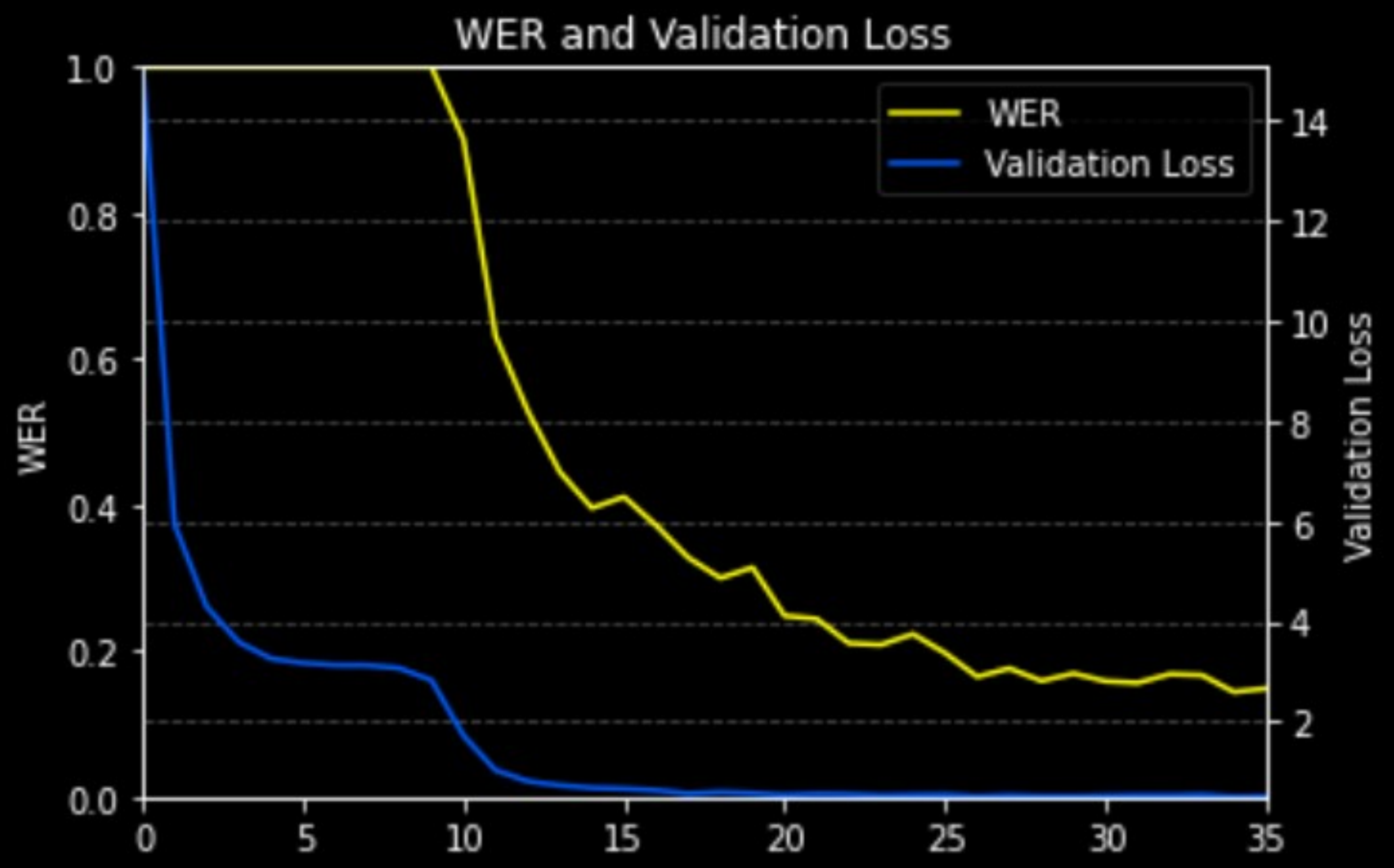


Experiment Settings

Two 1 hour long datasets were concatenated to give the data for training: The Phonetically Rich Urdu Speech (PRUS) corpus by CLE, and the Urdu subset of the Common-Voice 8.0 dataset by Mozilla. During fine-tuning, the feature extractor is frozen, and only the transformer is trained. The model is fine-tuned for 35 epochs.

Methodology

We used transfer learning to fine-tune Facebook's 300 M parameter multilingual XLS-R speech recognition model. The model consists of a CNN feature extractor that extracts latent speech representations from raw speech audio. A transformer then trains upon these representations by masked language modeling.



Results and Future

A word error rate of 17% was achieved. Since Facebook had pre-trained the XLS-R model on European languages, it gives poor results when finetuned on languages such as Urdu and Arabic. In future we plan to pretrain the monolingual word2vec2 model on Urdu data so that single digit error rates are possible.