Technical Disclosure Commons

Defensive Publications Series

November 2022

PRODUCTION-GRADE ONLINE SPEECH RECOGNITION FOR LOW-RESOURCE LANGUAGES

Sylvain Le Groux

Zili Huang

Follow this and additional works at: https://www.tdcommons.org/dpubs_series

Recommended Citation
Le Groux, Sylvain and Huang, Zili, "PRODUCTION-GRADE ONLINE SPEECH RECOGNITION FOR LOW-RESOURCE LANGUAGES", Technical Disclosure Commons, (November 14, 2022)
https://www.tdcommons.org/dpubs_series/5500

This work is licensed under a Creative Commons Attribution 4.0 License.
This Article is brought to you for free and open access by Technical Disclosure Commons. It has been accepted for inclusion in Defensive Publications Series by an authorized administrator of Technical Disclosure Commons.

PRODUCTION-GRADE ONLINE SPEECH RECOGNITION FOR LOW-RESOURCE LANGUAGES

AUTHORS:
Sylvain Le Groux
Zili Huang

ABSTRACT

Techniques are provided for speech recognition in real-time with a semi-supervised model based on Wav2vec2.0. Only minimal training data is required, thereby enabling service of under-represented/low resource languages at a quality level comparable to more widely available languages.

DETAILED DESCRIPTION

Most speech recognition systems that achieve an acceptable Word Error Rate (WER) are trained based on the English language, for which a significant amount of training data is available. Current state of the art techniques perform relatively poorly for other under-represented languages where annotated data is scarcer. Obtaining high-quality training data usually requires hand labeling and human curation, which is very costly. As a result, when investing in data acquisition, typical cost/benefit analysis favors the most spoken languages. As a result, there is a clear bias in speech recognition towards a few of the most spoken languages. Low-resource languages are under-represented and the corresponding speech recognizers have significantly lower quality. One way to mitigate this bias is to use semi-supervise learning methods that require minimal training data. However, in their standard form, these types of methods lead to inference algorithms that are ill-suited for real-time decoding, which is arguably the most relevant use-case for industry application (in the form of closed captions, for instance).

Despite great offline performance on real meeting recordings, the current state of the art wav2vec2.0 models, which combines self-supervised pre-training and supervised fine-tuning with Connectionist Temporal Classification (CTC) inference, does not support online ASR. Although CTC is online in nature, there are several designs in wav2vec 2.0 that make it difficult to perform online Automatic Speech Recognition (ASR).

In the wav2vec 2.0 architecture, the model takes the raw waveform and down-samples with a few convolution layers. After that, the extracted features may pass several transformer layers.

This architecture limits real-time inference for at least three reasons. First, with the self-attention mechanism in the transformer layers, each frame attends to the entire sequence. This makes online ASR impossible. Second, a 1D-convolution with kernel size 128 is used for positional embedding, which may involve examining 64 frames in the past and 64 frames in the future. This results in long latency for real-time applications. Third, group normalization in the convolutional layers compute statistics on the time axis, which makes online decoding impossible.

To enable real-time inference, a few modifications to the architecture are presented herein. For 1D-convolutions with large kernel sizes, the kernel size may be reduced from 128 to 16. Although the model still needs to look at some future context, the lookahead time is just 8 * 20 = 160ms which is short enough for real-time applications. Also, the group normalization may be removed from the convolution layers without losing significant accuracy of the ASR. Furthermore, an attention mask may be used, as illustrated in Figure 1 below. With this attention mask, the audio frames can only attend to the past or near futures, hence allowing for real-time decoding.

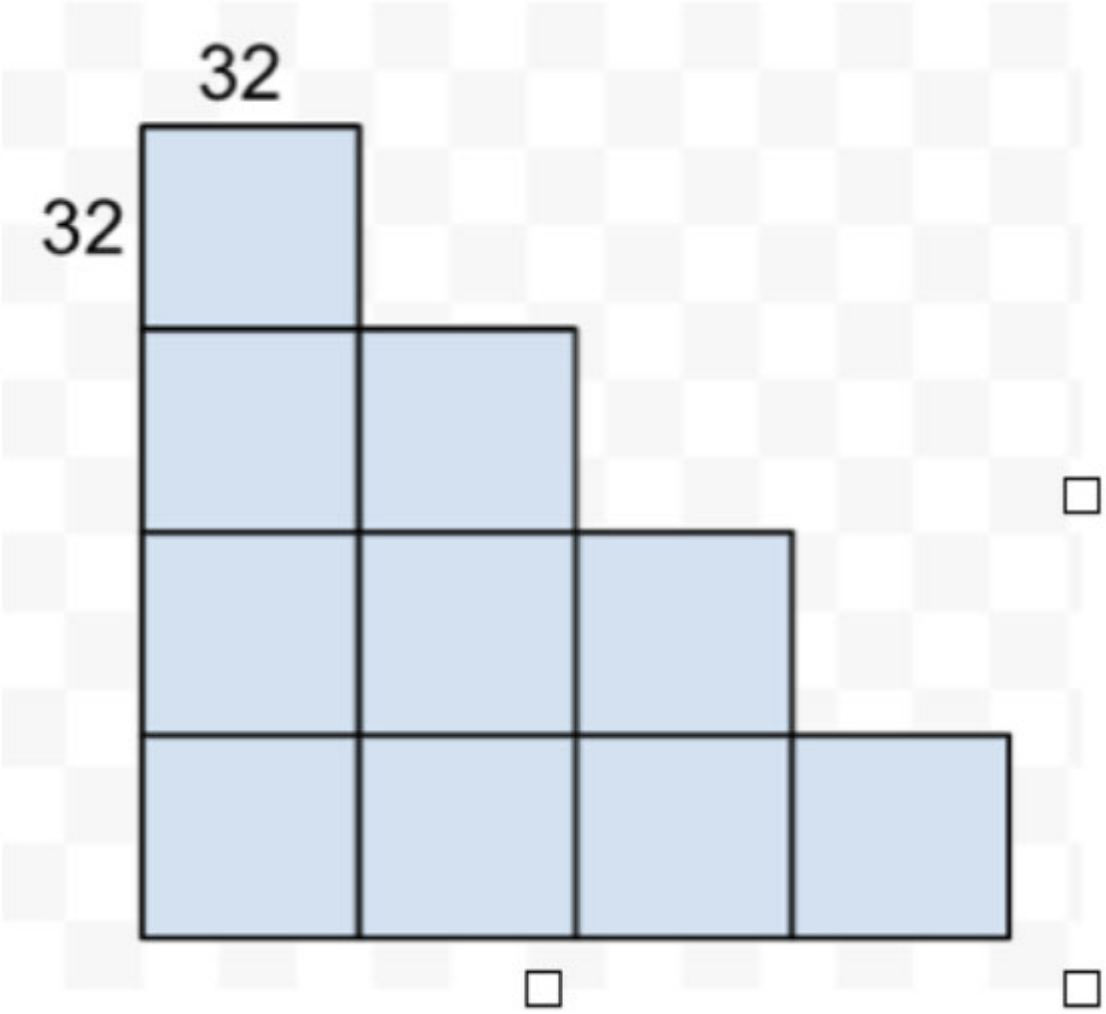

Figure 1

These changes enable real-time decoding for wav2vec 2.0 while preserving its impressive WER performance. The online results are shown in Table 1 below.

system	offline / online	WER (%) test clean	WER (%) test other
original	offline	6.1	13.4
chunk 32 (0.64s)	online	7.9	21.2
chunk 64 (1.28s)	online	7.5	19.6
chunk 128 (2.56s)	online	6.8	17.6

Table 1

As shown, the online system performs better with larger chunk sizes. When the chunk size is 128 (2.56s), the WER is 6.8% on LibriSpeech test clean (0.7% worse than offline model) and 17.6% on LibriSpeech test other (4.2% worse than offline model).

In summary, techniques are provided for speech recognition in real-time with a semi-supervised model based on Wav2vec2.0. Only minimal training data is required, thereby enabling service of under-represented/low resource languages at a quality level comparable to more widely available languages.