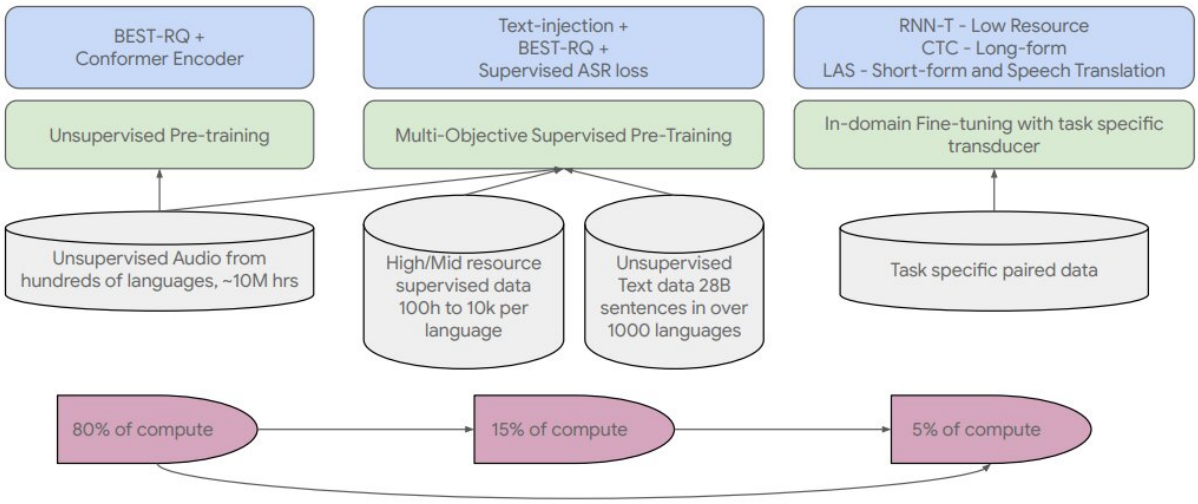
Google gives progress report on its Universal Speech Model

March 7 2023, by Bob Yirka



An overview of our approach. Training is split into three stages. (i) The first stage trains a conformer backbone on a large unlabeled speech dataset, optimizing for the BEST-RQ objective. (ii) We continue training this speech representation learning model while optimizing for multiple objectives, the BEST-RQ objective on unlabeled speech, the modality matching, supervised ASR and duration modeling losses on paired speech and transcript data and the text reconstruction objective with an RNN-T decoder on unlabeled text. (iii) The third stage fine-tunes this pre-trained encoder on the ASR or AST tasks. Credit: *arXiv* (2023). DOI: 10.48550/arxiv.2303.01037

In November, Google announced that it was embarking on an initiative that would culminate in the development of a machine-learning model

capable of recognizing and translating 1,000 of the world's most spoken languages. Over the past several months, the company has been working toward that goal and has published a blog entry by members of the team working on the project. The team at Google has also published a paper describing the introduction of its Universal Speech Model (USM) on the *arXiv* pre-print server.

The updates provided by Google are part of a more overarching goal: to create a language translator using automatic speech recognition (ASR) capable of translating any language in the world on demand. To that end, they have chosen to temporarily cap the number of languages they are attempting to support (at 100) due to the low numbers of people who speak less common languages. Such rare languages lack datasets for training.

As part of their announcement, Google outlined the first steps toward their USM—breaking it down into families of speech models trained on billions of hours of recorded speech and spanning over 300 languages. They note that their USM is already currently used for closed-captioned language translations on YouTube. They also outline the generic model for each of the families.

Google explains that the models are being produced using training "pipelines" that involve three kinds of datasets: unpaired audio, unpaired text and paired ASR data. They also note that they are using conformer models to handle the expected 2B parameters required for the project and will do so using three major steps: unsupervised pre-training, multi-objective supervised pre-training and supervised ASR training. The end result will be the production of two types of models—those that are pretrained and ASR models.

Google further claims that in its current state, its USM has shown comparable or superior performance to the Whisper model—a general-

purpose speech recognition model created by the GitHub community. In addition to using the USM for YouTube, Google is expected to pair its model with other AI applications, including augmented reality devices.

More information: Yu Zhang et al, Google USM: Scaling Automatic Speech Recognition Beyond 100 Languages, *arXiv* (2023). DOI: 10.48550/arxiv.2303.01037

© 2023 Science X Network

Citation: Google gives progress report on its Universal Speech Model (2023, March 7) retrieved 20 July 2024 from https://techxplore.com/news/2023-03-google-universal-speech.html

This document is subject to copyright. Apart from any fair dealing for the purpose of private study or research, no part may be reproduced without the written permission. The content is provided for information purposes only.