1000 Words in 1000 Languages

Vijay Janapa Reddi, Pete Warden, Greg Diamos
Harvard, Google, Landing.AI
MLCommons

Objective

Enormous effort has been spent to collect and label millions of examples in datasets such as ImageNet, COCO, or OpenImages. For each new machine learning task, these large datasets are often only used in the preliminary step of pretraining weights on a related task, and a second dataset must then be manually collected and labeled in order to train on the actual desired objective. This additional dataset curation is time-consuming (i.e., slow time-to-solution) and expensive.

In “1000 Words in 1000 Languages,” we aim to provide a suite of data engineering techniques that automate dataset generation for embedded machine learning tasks, without the need for manual data collection and labeling. Our goals are twofold. First, we are building a data engineering pipeline as a community driven research project to enable automatic collection and labeling for commonly spoken words in languages that are widely spoken like English. Second, we seek to expand the scope to include more diverse languages with the steadfast goal of expanding embedded tinyML benefits to all.

Rapid Data-to-Model Iteration

Our candidate example task is embedded keyword-spotting (KWS), where a small, low-power model continually listens for specific keywords in speech (e.g. “OK Google” or “Alexa”), for instance in order to activate a voice assistant. By convention, training a new KWS model would necessitate the manual collection of thousands of examples of labeled audio clips for each keyword. Instead, we repurpose already-labeled data and trained models for an adjacent non-embedded task, general speech recognition. Using Mozilla’s CommonVoice, a public multilingual crowdsourced dataset, along with available general speech-to-text (STT) models we automate dataset generation for custom embedded KWS models. Since these general speech-to-text models provide metadata features containing approximate timings for word utterances, we can automatically extract desired keywords from CommonVoice’s large corpus of unaligned sentence recordings. Once enough keywords are extracted, we seek to automatically train and evaluate high-quality embedded KWS models.

Even with noisy precision and recall in a STT model, we believe the approach is extensible to leveraging unlabeled speech found “in the wild” such as from online videos with the appropriate licensing. A user can therefore simply provide a list of perhaps a dozen desired keywords they wish to detect, and our pipeline will automate the extraction, training, and validation of a model without requiring any labeling. For “long-tail” infrequently used words, we can reduce the barrier to entry for a custom model by providing a candidate dataset to the user with noisy labels which only need verification, removing the burden of data collection.

Join us!

We believe this approach will enable much wider access to machine learning and automation for users who do not have the resources to manually label and train models themselves. By combining large amounts of labeled and unlabeled data with noisy inferences from domain-adjacent pretrained models, we aim to provide a “self-serve” automated pipeline for model training on custom embedded tasks. If you have ideas on what you think we should tackle next, reach out and let us know!

Team Members

Mark Mazumder (Harvard), Tejas Prabhune (Harvard), Colby Banbury (Harvard), Pete Warden (Google), Vijay Janapa-Reddi (Harvard), Greg Diamos (Landing.AI) -- Join us and democratize data engineering pipelines for the people!