Progress Update

Yu-Hsiang Lin
Language Technologies Institute
Carnegie Mellon University
yuhsianl@andrew.cmu.edu

Abstract

Place holder [1].

1 Question to answer

How can we create a speech recognition system that transcribes low-resource languages into International Phonetic Alphabets (IPAs) as accurately and quickly as possible?

2 Naive way to proceed

Based on [3].

1. Prepare data: N high-resource languages that have transcription from speech to IPAs (or speech $\rightarrow$ language $\rightarrow$ IPA?), and a target low-resource language with limited transcription.
2. Train a multi-lingual ASR (N high-resource languages $\rightarrow$ IPA) model: CTC objective function $+$ biLSTM.

   (Model parameters $=$ LSTM only? Input acoustic representation is engineered. Output phonemic sequence is computed by CTC decoding.)

   Can be either warm-start (training data contains target low-resource languages) or cold-start (training data contains only high-resource languages).
3. Given a target low-resource language, incrementally train (fine-tune) the multi-lingual model: 3 possibilities $\Rightarrow$
   (a) Train on only the target low-resource language.
   (b) Select a high-resource language (that is used previously in the multi-lingual model training) that is similar to the target low-resource language, and train on their data combined (concatenation for accuracy, or sampling for speed).
   (c) Think of some way to use the untranscribed speech recording (could actually be plenty of them) to help the training on the target low-resource language.
   (May be an independent topic outside the multi-lingual setting?)

3 Other possibilities

- Meta learning [2]: learn a good initialization.
- Some way to use the untranscribed speech recording to help the training on the target low-resource language (without multi-lingual pre-trained model).

  Will the (plenty of) untranscribed data help regularization/generalization?

  $\Rightarrow$ An idea: The typical pattern of semi-supervised learning is (similar to EM): Do iterations. Within each iteration we have two stages. First stage: transcribe the untranscribed data using the current model. Second stage: either maximize the expectation of the model of predicting such transcription (EM), or compute the similarity between the unlabeled speech and the labeled speech with the same predicted labels, and penalize the diversity.

Question is: how to update the model (at least in the second way)? How can I use the signal coming from the diversity to update my model?

- Cotraining? Separate features into two groups and train two ASR models?

References

[1] Oliver Adams, Trevor Cohn, Graham Neubig, and Alexis Michaud. Phonemic transcription of low-resource tonal languages. In *Australasian Language Technology Association Workshop 2017*, December 2017.

[2] Jiatao Gu, Yong Wang, Yun Chen, Kyunghyun Cho, and Victor O.K. Li. Meta-learning for low-resource neural machine translation. In *EMNLP*, 2018.

[3] Graham Neubig and Junjie Hu. Rapid adaptation of neural machine translation to new languages. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Brussels, Belgium, November 2018.