Contextual Speech Recognition with Difficult Negative Training Examples

Uri Alon, Golan Pundak, Tara N. Sainath

uralon@cs.technion.ac.il  {golan,tsainath}@google.com

1. Task: Contextualized ASR
- Context provided in addition to audio can help reduce WER significantly.
- Such user-specific contextual information can include:
  - The user’s list of songs
  - The user’s contact list
  - The currently installed apps
- Proper nouns are very frequent in various ASR tasks:
  - “Call Joan’s mobile”
  - “Play Taylor Swift”
  - “How tall is LeBron James?”
- But contextual ASR models usually perform poorly on rare words and especially on proper nouns (NNPs).

2. The Contextualized LAS (CLAS) Model (Pundak et al., SLT’2018)
- CLAS is an E2E ASR model based on the Listen-Attend-and-Spell (LAS) encoder-decoder architecture.
- The key difference from LAS: biasing sub-module.

3. The Problem: The Network Fails to Distinguish Between Phonetically Similar Phrases
- Disambiguation of similarly sounding phrases is challenging.
- The network makes even more mistakes as the set of bias phrases becomes larger.

4. Training with Difficult Negative Examples
- During training, we provide the network with phonetically similar proper nouns (NNPs) as the “distractors”.
- This way, we encourage the network to:
  - Distinguish between similarly sounding phrases
  - Learn more discriminative representations.

5. Evaluation
- We experimented with the following training schemes:
  - Vanilla CLAS
  - CLAS+NNP
  - CLAS+fuzzy
  - CLAS NNP+fuzzy
- Bias Phrases Selection: Random
- Distractions Selection: Random NNPs
- Results:
  - Table: WER of the compared training schemes. In parentheses: the relative improvement over Vanilla CLAS.

6. Qualitative Analysis
- The fuzzy model tends mostly to “creepy carrots” and makes a correct prediction, while the non-fuzzy model tends to “sleepy carrots” and predicts the wrong word “sleepy.”