Overview

- Background
  - Today's automatic speech recognition (ASR) systems heavily rely on supervised training using large amounts of task-matched training data.
  - The cost of transcribing speech data is repeatedly required to support new languages and new tasks.
  - A system would become more self-sufficient and useful if it possessed the ability to learn from very light feedback from the users.
- Our contribution
  - Formulate a general reinforcement learning framework for ASR systems based on the policy gradient method.
  - Propose a hypothesis selection method following the reinforcement learning framework, where the feedback is given by the user selection of hypotheses.

Design of user feedback

- Accuracy-based feedback
  - Calculating word accuracy is difficult and time consuming for the user.
- Selection-based feedback (Proposed method)
  - Two recognition systems present hypotheses to the user.
  - The user selects the best hypothesis among them.

Implementation with Approximation

- Hypothesis generation: Sampling from posterior distribution → Viterbi decoding.
- Rival system
  - Use the \( n \)-th \((1 \leq n)\) best hypothesis of the same system as the rival hypothesis.
  - Hypothesis \( (tg) \): The Candidate 1 hypothesis \( f(1) \)
  - Hypothesis \( (rv) \): The Candidate 2 hypothesis \( f(2) \)
- Parameter update: Utterance based update
  - Weighted gradient
    \[
    (1 + a) \left( \frac{1}{1 + a} \right) \frac{\partial \log P_{AM}(l, | s_t)}{\partial \theta} \quad \text{Candidate 1}
    \]
    \[
    (1 + a) \left( 1 - \frac{1}{1 + a} \right) \frac{\partial \log P_{AM}(l, | s_t)}{\partial \theta} \quad \text{Candidate 2}
    \]
    \[
    \alpha \geq 1 \quad \alpha = 1 \rightarrow \text{Large batch based update}
    \]
    \[
    \alpha = 0 \rightarrow \text{Gradient ascent based solution}
    \]
- General form of REINFORCE algorithm [Williams, 1992]
  - \( (r \neq b) \frac{\partial \log \theta}{\partial \theta} = \theta \): Neural network based policy function
  - \( \theta \): Parameters of the neural network

Results (without Hypotheses Selection Error)

- Number of stages and WERs of the large batch data
  - The proposed method reduced WER compared to the initial model.
  - Future work: Improving the stability to over-training and the learning efficiency for the user feedback.

Conclusions

- Formulated a policy gradient-based reinforcement learning framework for ASR systems, and proposed a hypothesis selecting-based reinforcement learning method.
- The proposed method reduced WER compared to the unsupervised adaptations.
- Future work: Improving the stability to over-training and the learning efficiency for the user feedback.