On the Accuracy and Inference Efficiency for Low-resource Automatic Speech Recognition

Ruchao Fan
Advisor: Prof. Abeer Alwan
UCLA Speech Processing and Auditory Perception Lab
Automatic speech recognition (ASR)

Conventional v.s. End-to-end Model

Data hungry!

The figure is borrowed from Dr. Sainath talk @Samsung AI Forum
End-to-end Models for Low-resource ASR

- End-to-end Models
  - **Pros:**
    - Joint optimization of modules and reduce intermediate errors
    - A simplified pipeline that is easy for training and deployment
  - **Cons:**
    - Paired audio-annotation data for training -> Need more data!
    - Autoregressive generation of the output sequence -> computational inefficient!

- Low-resource ASR
  - Limited annotated data, i.e. child speech, privacy issues -> Low Accuracy
  - Using autoregressive end-to-end models -> Low efficiency
My research - child speech recognition

• Improving accuracy
  • The performance of ASR systems for child speech is inadequate for practical usage, e.g. more than 30% word error rate (WER) in comparison
  • Solution:
    • Self-supervised learning for unsupervised pretraining by using easy-accessible untranscribed data

• Improving inference efficiency
  • Efficient inference benefits on-device deployment
  • Solution:
    • Non-autoregressive framework for end-to-end speech recognition
Part I. Self-supervised learning for unsupervised pretraining
Previous work

- Bidirectional autoregressive predictive coding (Bi-APC) for unsupervised pre-training and its application to children’s ASR

- **Goal:** Develop pre-training methods for improving children’s ASR performance using adult speech data.

Model pre-training

- Two-step process:
  - Pre-training on a data-sufficient task (adult acoustic models)
  - Fine-tuning on the target low-resource task (child acoustic models)

- Pre-training methods depending on whether the pre-training data is labelled:
  - Supervised pre-training (SPT)
  - Unsupervised pre-training (UPT)
Proposed Bidirectional APC (Bi-APC)

- **Motivation:** Bidirectional models, like BLSTM outperform their unidirectional counterparts for ASR, and APC is not suitable for BLSTM.
- **Proposed Bi-APC:** Decompose forward computation of BLSTM into
  - **Forward path:** predict a frame \( n \) steps after the current frame given all the **past frames**.
  - **Reversed path:** predict a frame \( n \) steps before the current frame given all the **future frames**.
Part II - Single-step Non-autoregressive Transformer
Previous work

- CTC alignment-based single step non-autoregressive transformer for speech recognition
- CTC: connectionist temporal classification

- **Goal:** Achieve faster generation of output sequence with as less performance degradation as possible

Recall
Autoregressive vs. Non-autoregressive

1. Autoregressive (AR) transformers generate tokens step by step

2. Non-autoregressive transformers (NAT) achieve parallel generation
Proposed CASS-NAT

- The idea is to replace word embedding with the token-level acoustic embedding.
Token-level acoustic embeddings express both the acoustic and semantic similarities.

The behaviour of containing semantic information is similar to word embedding, which explains why CASS-NAT works as well as AT.
Thank you for listening!

Q & A!
Email: fanruchao@g.ucla.edu