Language Model Integration Based on Memory Control for Sequence to Sequence Speech Recognition

Jaejin Cho¹, Shinji Watanabe¹, Takaaki Hori², Murali Karthick Baskar³, Hirofumi Inaguma⁴, Jesus Villalba¹, Najim Dehak²

¹Center for Language and Speech Processing, Johns Hopkins University
²Mitsubishi Electric Research Laboratories (MERL)
³Brno University of Technology
⁴Kyoto University

Introduction
- Task: Language model (LM) integration to help sequence-to-sequence (S2S) ASR training
- Proposal:
  - Update of the hidden/cell states in S2S LSTM decoder using LM information
  - Use of the LM information for both character inference and states update in decoder
  - 3 variants with the idea
- System:
  - S2S attention model with CTC-loss as a regularizer
  - LM trained ahead of the S2S model training

Background: LM integration in S2S model
- LM integration in decoding:
  - Shallow fusion (SF): Linear interpolation between two scores with a hyper parameter
  \[ \hat{y} = \arg\max \left( \log p(y|x) + \gamma \log p(y) \right) \]
  - Deep fusion (DF): Parameter learning to connect LM and S2S model

Proposed method: Memory control fusion (MF)
- Belongs to the second category, LM integration in training
- Controls the hidden/cell (memory) states in S2S decoder using LM information
- Affects both inference and the states update in the decoder over time

Experimental setup
- Mono-lingual ASR
  - Paired data (Speech and its transcript): Librispeech 100/960 hrs
  - External text (not paired with speech): 10 times of whole paired text
  - S2S: 8-layer BLSTM encoder + 1-layer LSTM decoder, LM: 2-layer LSTM
- Transfer learning from a language-independent model to a target model
  - Language-independent model: Trained with 10 Babel languages (~643 hrs) not including the target language
  - Target model: Trained on a target language data, Swahili (~50 hrs) with initialized parameters from the language-independent model

Results
- Mono-lingual setup
  - Consistent improvements compared to the previous methods
  - Third variant updating both hidden/cell states worked the best
  - ~2 to 4% relative improvement in WER in mono-lingual setup
  - ~9 to 10% relative improvement in WER in multi-lingual transfer learning setup

Conclusion

Acknowledgements

The work reported here was started during JSALT 2018, and supported by JHU with gifts from Amazon, Facebook, Google, Microsoft and Mitsubishi Electric.