16 Learning speech recognition from small paired data

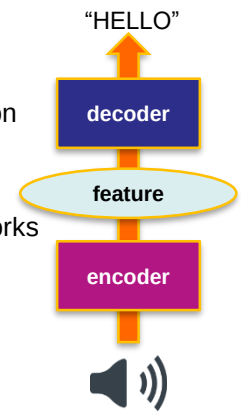
- Semi-supervised end-to-end training with text-to-speech -

Abstract

We propose a semi-supervised end-to-end method for learning speech recognition from small paired data and large unpaired data. This is because preparing the paired data of a speech and its transcription text requires a large amount of human effort. In our method, we introduce speech and text autoencoders that share encoders and decoders with an automatic speech recognition (ASR) model to improve ASR performance using speech-only and text-only training datasets. To build the speech and text autoencoders, we leverage state-of-the-art ASR and text-to-speech (TTS) encoder-decoder architectures. These autoencoders learn features from speech-only and text-only datasets by switching the encoders and decoders used in the ASR and TTS models. Simultaneously, they aim to encode features to be compatible with ASR and TTS models using a multi-task loss.

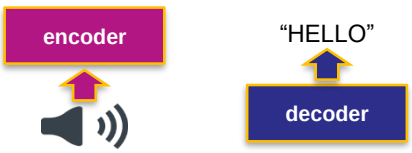
Speech Recognition

- Prepare paired data of speech and transcription text for training
- Train speech encoder and text decoder networks
- Perform speech recognition on given speech input



Problems

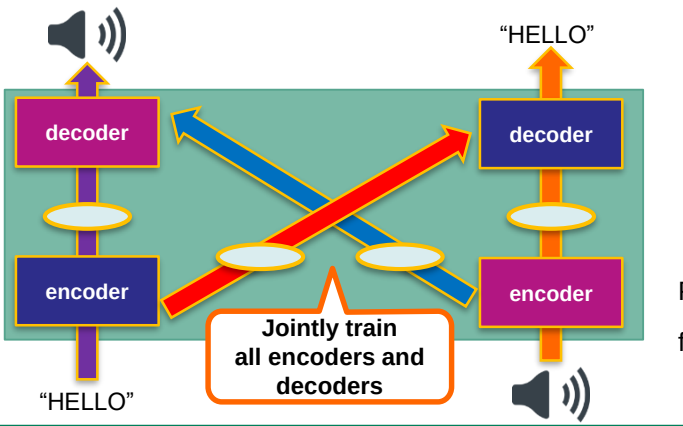
- The encoder-decoder model requires a large amount of the paired speech and transcription text dataset.
- Preparation of such dataset needs huge amounts of time and money
- If the networks can learn speech-only and text-only datasets, the data preparation becomes much easier



Semi-supervised training method

Point 1: Combine speech-to-text task with text-to-speech

- Train with paired data: speech-to-text, text-to-speech
- Train without paired data: speech-to-speech, text-to-text



Training task				
speech to text	text to text	speech to speech	text to speech	char error rate
✔				15.0 %
✔	✔			9.0%
✔		✔		8.7%
✔	✔	✔	✔	**8.4%**

Point2: Training to encode or decode features that look similar to each other

References

[1] S. Karita, S. Watanabe, T. Iwata, A. Ogawa, M. Delcroix, "Semi-supervised end-to-end speech recognition," in Proc. of *2018 Interspeech*, pp. 2-6, 2018.

[2] S. Karita, S. Watanabe, T. Iwata, M. Delcroix, A. Ogawa, T. Nakatani, "Semi-supervised end-to-end speech recognition using text-to-speech and autoencoders," in Proc. of *2019 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2019.

Contact

Shigeki Karita Email: cs-liaison-ml at hco.ntt.co.jp
Signal Processing Research Group, Media Information Laboratory

Innovative R&D by NTT
Open House 2019