EXPLORING TRANSFER LEARNING FOR E2E SPOKEN LANGUAGE UNDERSTANDING

Subendhu Rongali*, Beiye Liu, Liwei Cai,
Konstantine Arkoudas, Chengwei Su, Wael Hamza
Alexa AI New York, *University of Massachusetts Amherst

PROBLEM STATEMENT

ASR

play watermelon sugar

NLU

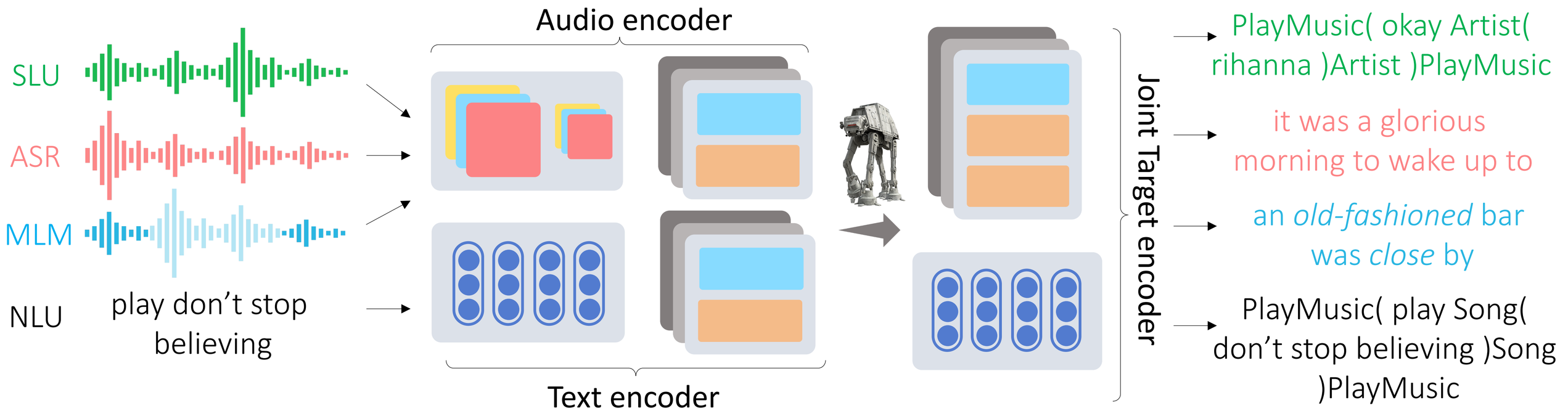
PlayMusic(play **Song(** watermelon sugar **)Song)PlayMusic**

E2E

ADVANTAGES
1. Smaller, faster engine
2. No cascading errors
3. Possibly better optimization

ISSUES
1. Require huge amts of E2E data
2. Lots of ASR, NLU data not used
3. Feature Expansion not easy

OUR APPROACH: AT-AT (AUDIO-TEXT ALL-TASK) TRANSFORMER



KEY POINTS

- Audio is processed as log-filterbank fts (80 dim).
- Audio encoder consists of a 2L 2D conv net embedder and a 12L transformer encoder.
- Text encoder is BERT-base-uncased (768 dim, 12L).
- Decoder consists of a tied target embedding-generator block, and a 6L transformer decoder.
- Label smoothing, Noam scheduler, Beam decoding.

RESULTS

Model	ER	EM Acc
Alexa Music Data		
E2E Model (10% data)	27.86	55.99
E2E Model (100% data)	19.23	67.81
AT-AT (10% + Music ASR)	**20.68**	**66.32**
AT-AT (100% + LibriSpeech ASR)	**18.07**	**69.20**
FluentSpeech		
SOTA [Lugosch et al. 2019]	1.2	98.8
E2E Model	9.3	91.7
AT-AT (+LibriSpeech/Music data)	**0.5**	**99.5**
Facebook TOP (Zeroshot)		
E2E Model (synthetic audio)	--	69.19
AT-AT (only text)	**--**	**51.54**
AT-AT (text + synthetic audio)	**--**	**70.60**

DISCUSSION

- The task is passed as <BOS> token for decoding.
- AT-AT is similar in spirit to Google's T5 [Raffel et al. 2019], a Seq2Seq text-text multi-task model.
- We beat an E2E model trained on 10% data by using other ASR transcriptions., and an E2E model trained on 100% data by using data from LibriSpeech.
- We improve upon the SOTA performance on the external FluentSpeech and SNIPS datasets.
- These improvements can be attributed to our model learning a better distribution over the English language during ASR.
- AT-AT can perform zeroshot E2E learning both with and without access to a TTS system. It achieves SOTA results on the Facebook TOP dataset.

REFERENCES

1. A Comparative Study on Transformer vs RNN in Speech Applications. Karita et al 2019. arXiv:1909.06317v2 [cs.CL]
2. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. Raffel et al. 2019. arXiv:1910.10683v2 [cs.LG]
3. Speech Model Pre-training for End-to-End Spoken Language Understanding. Lugosch et al. 2019. arXiv:1904.03670v2 [eess.AS]