Woojin Choi

[1] NeMo : a toolkit for conversational AI, https://github.com/NVIDIA/NeMo

Topic : English & Korean Automatic Speech Recognition for Understanding Voice-Commands

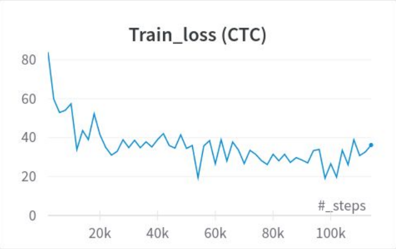
[Current Task]

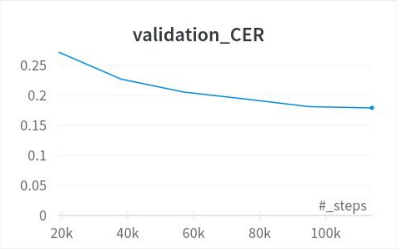
- Train ASR model w/ Korean & English speech dataset

```
# EXAMPLE : "Kspon_manifest.json"
# audio_filepath                duration                text
{"audio_filepath": "./KsponSpeech_620001.wav", "duration": 3.527, "text": "그러면 너랑 아 뭐 카페
가자는 거야?"}
..
..
..
{"audio_filepath": "./KsponSpeech_620004.wav", "duration": 3.414, "text": "정말 좋은
일이로구나."}
{"audio_filepath": "./KsponSpeech_620005.wav", "duration": 2.434, "text": "한 번 물어보면 안
돼?"}
```

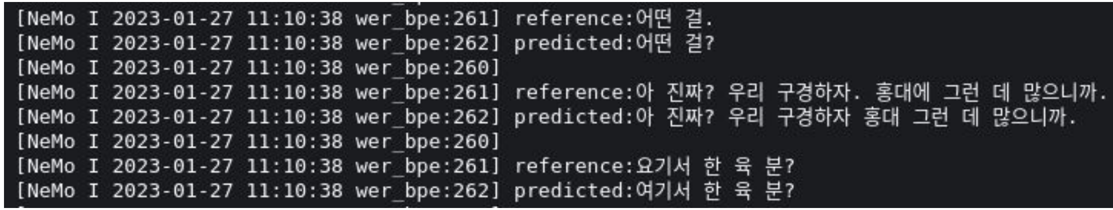
[This Week]

- **Nvidia NeMo**
- Training my NeMo model
  1) Dataset : KsponSpeech (Korean, ~1000 hrs)
  2) Preprocessing : file (.pcm -> .wav), cleanup text
  3) Pre-trained model : Conformer-CTC(small) - English, 13.9M parameters
  4) Training config : Subword tokenizer (BPE), batch=32, epochs=10 (~6)
- Results
  - Time / Memory : 3~ (hrs/epoch) / ~17 (GB) -> *Multi-GPU, increase batch*
  - Evaluation : ~~WER~~ / **CER = 0.18 (character-error rate)**





- Further experiments
  - Tokenizer type, vocab-size (=5000)
  - Model type : CTC, Transducer (loss)
  - Training config : batch-size, epochs, optimizer, etc.
  - English model : pre-trained vs. trained w/ dataset (Google Speech Commands)
  - *Inference w/ mic-input*

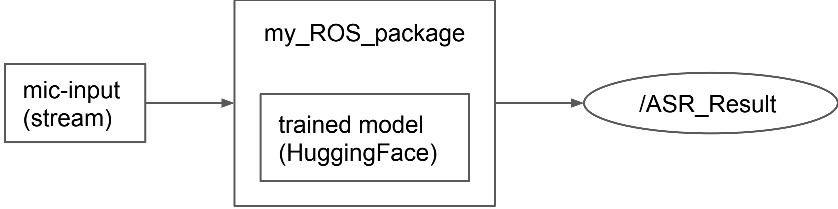


```
[NeMo I 2023-01-27 11:10:38 wer_bpe:261] reference:어떤 걸.
[NeMo I 2023-01-27 11:10:38 wer_bpe:262] predicted:어떤 걸?
[NeMo I 2023-01-27 11:10:38 wer_bpe:260]
[NeMo I 2023-01-27 11:10:38 wer_bpe:261] reference:아 진짜? 우리 구경하자. 홍대에 그런 데 많으니까.
[NeMo I 2023-01-27 11:10:38 wer_bpe:262] predicted:아 진짜? 우리 구경하자 홍대 그런 데 많으니까.
[NeMo I 2023-01-27 11:10:38 wer_bpe:260]
[NeMo I 2023-01-27 11:10:38 wer_bpe:261] reference:요기서 한 육 분?
[NeMo I 2023-01-27 11:10:38 wer_bpe:262] predicted:여기서 한 육 분?
```

[Next Week]

1) Training & fine-tuning
2) Write voice-command recognition program
3) Study ROS : python -> package

```
# EXAMPLE : "Kspon_manifest.json"
# audio_filepath                duration            text
{"audio_filepath": "./KsponSpeech_620001.wav", "duration": 3.527, "text": "그러면 너랑 아 뭐 카페
가자는 거야?"}
..
..
..
{"audio_filepath": "./KsponSpeech_620004.wav", "duration": 3.414, "text": "정말 좋은
일이로구나."}
{"audio_filepath": "./KsponSpeech_620005.wav", "duration": 2.434, "text": "한 번 물어보면 안
돼?"}
```

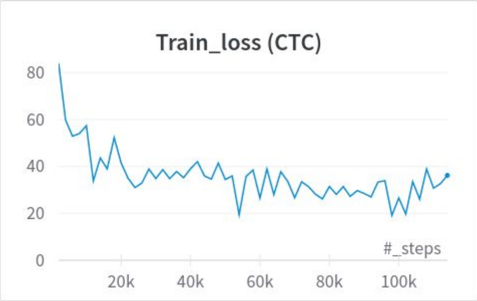


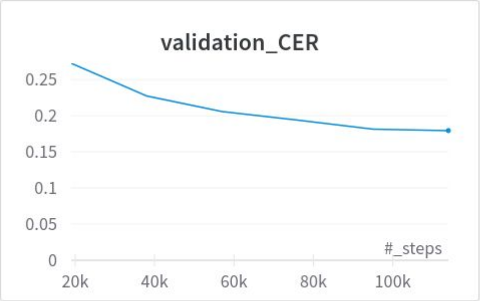
mic-input (stream)

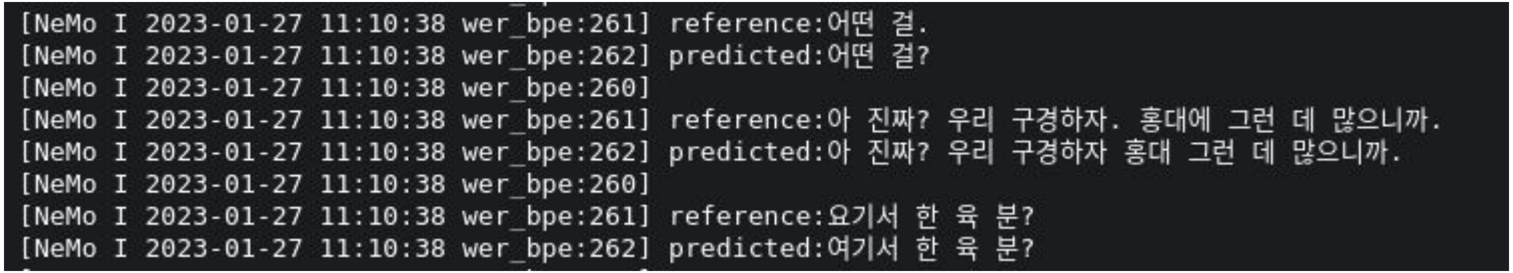
my_ROS_package

trained model (HuggingFace)

/ASR_Result







```
[NeMo I 2023-01-27 11:10:38 wer_bpe:261] reference:어떤 걸.
[NeMo I 2023-01-27 11:10:38 wer_bpe:262] predicted:어떤 걸?
[NeMo I 2023-01-27 11:10:38 wer_bpe:260]
[NeMo I 2023-01-27 11:10:38 wer_bpe:261] reference:아 진짜? 우리 구경하자. 홍대에 그런 데 많으니까.
[NeMo I 2023-01-27 11:10:38 wer_bpe:262] predicted:아 진짜? 우리 구경하자 홍대 그런 데 많으니까.
[NeMo I 2023-01-27 11:10:38 wer_bpe:260]
[NeMo I 2023-01-27 11:10:38 wer_bpe:261] reference:요기서 한 육 분?
[NeMo I 2023-01-27 11:10:38 wer_bpe:262] predicted:여기서 한 육 분?
```

figures

1. Nemo .json file example
2. log - predictions
3. W&B - train/val loss graph
4. diagram - my ROS package