fairseq S2T: Fast Speech-to-Text Modeling with fairseq

12/6/2020 @ AACL 2020

Changhan Wang, Yun Tang, Xutai Ma, Anne Wu, Dmytro Okhonko, Juan Pino

FACEBOOK AI

Overview

- Fairseq: PyTorch-based sequence modeling framework
  - WMT-winning MT models, RoBERTa, XLM-R, BART, etc.
- Fairseq S2T: an extension for speech-to-text tasks (S2T)
  - Speech recognition (ASR), speech translation (ST), etc.
  - Natural language modality: {text} → {text, speech}
  - Design goals: integration, scalability & reproducibility

Design

- Integration
  - All-in-one: ASR, ST, language modeling (LM), machine translation (MT), non-autoregressive MT, online MT/ST, self-supervised pre-training, etc.
  - End-to-end workflows from speech processing to model evaluation
- Scalability
  - Multi-node, mixed precision, model parallelism, speech data I/O
- Reproducibility
  - Training recipes, checkpoints

	ASR	LM	MT	Non-Autoreg. MT	Offline ST	Online ST	Speech Pre-training	Multi-node training	Pre-trained models
ESPNet-ST	✓	✓	✓		✓			✓[†]	✓
Lingvo	✓	✓	✓		✓[‡]			✓	
OpenSeq2seq[1]	✓	✓	✓					✓	✓
RETURNN[2]	✓	✓	✓		✓			✓	✓
SLT.KIT[3]	✓		✓		✓				✓
Tensor2Tensor[4]	✓	✓	✓					✓	✓
OpenNMT[5]	✓	✓	✓					✓	✓
Kaldi[6]	✓	✓							✓
Wav2letter++[7]	✓	✓							✓
fairseq S2T	✓	✓	✓	✓	✓	✓	✓	✓	✓

Table 1: Comparison of FAIRSEQ S2T with counterpart toolkits (as of July 2020). [†] Only available in version 2 (under development). [‡] Not publicly available. [1] Kuchaiev et al. (2018). [2] Zeyer et al. (2018). [3] Zenkel et al. (2018). [4] Vaswani et al. (2018). [5] Klein et al. (2017). [6] Povey et al. (2011). [7] Pratap et al. (2018).

(Comparison with counterparts)

Features

- Tasks: CTC/cross-entropy criterion, multilingual training with data resampling
- Models: RNN-based / Transformer-based
- End-to-end workflow
  - Online/offline speech feature extraction & transforms
  - Online tokenization, segmentation sampling (e.g. unigram)
  - Kaldi-like manifest TSV, configuration YAML, etc.
  - Metrics: WER, BLEU, chrF; AL, DAL (for online models)
  - Visualization: integration of TensorBoard & VizSeq

ASR Benchmark: LibriSpeech

- 1000h English speech
- Various size of Transformer models (T-Sm, T-Md & T-Lg) and RNN-based model (B-Big) trained with cross entropy criterion
- Competitive results with default model hyper-parameters and LR schedule

	Dev		Test	
	Clean	Other	Clean	Other
LAS†	-	-	2.8	6.8
Transformer‡	2.5	6.7	2.9	7.0
B-Big	3.7	11.4	3.9	11.5
T-Sm	4.1	9.3	4.4	9.2
T-Md	3.5	8.1	3.7	8.1
T-Lg	3.3	7.7	3.5	7.8

Table 4: FAIRSEQ S2T models on LibriSpeech (using default hyper-parameters and LR schedule). Dev and test WER reported. † Park et al. (2019). ‡ Synnaeve et al. (2019).

ST Benchmark: MuST-C

- 8-language translations for 400-500h English speech
- Bilingual: comparable to SOTA with similar model capacity (T-Sm), but with fewer tricks (speed perturbation, etc.)
- Multilingual: SOTA results with a larger model (T-Md)
- Simultaneous translation: different levels of latency-quality trade-offs (high/mid/low)

		De	Nl	Es	Fr	It	Pt	Ro	Ru
	Transformer[1]	17.3	18.8	20.8	26.9	16.8	20.1	16.5	10.5
	Transformer[2†]	22.9	27.4	28.0	32.7	23.8	28.0	21.9	15.8
	T-Sm	22.7	27.3	27.2	32.9	22.7	28.1	21.9	15.3
	Multi. T-Md*	24.5	28.6	28.2	34.9	24.6	31.1	23.8	16.0
B-Base	Offline	19.2	23.5	24.0	29.1	16.4	23.5	19.7	13.7
B-Base	High Lat.‡	18.6 (6.8)	22.9 (6.9)	22.3 (6.8)	28.4 (6.7)	15.4 (6.8)	22.6 (6.9)	19.1 (6.7)	12.9 (6.9)
B-Base	Mid Lat.‡	14.1 (5.4)	17.9 (5.4)	17.2 (5.5)	25.0 (5.3)	12.0 (5.5)	17.7 (5.8)	15.0 (5.6)	7.2 (5.8)
B-Base	Low Lat.‡	8.2 (2.9)	12.3 (2.8)	13.0 (3.0)	21.1 (2.8)	6.7 (2.9)	13.3 (2.9)	12.1 (2.9)	4.9 (2.7)

Table 2: FAIRSEQ S2T models on MuST-C. Test BLEU reported (for online models, AL is shown in parentheses). [1] Di Gangi et al. (2019). [2] Inaguma et al. (2020). † Applied additional techniques: speed perturbation, pre-trained decoder from MT and auxiliary CTC loss for ASR pre-training. ‡ Online models using beam size of 1 (instead of 5). * Trained jointly on all 8 languages.

ST Benchmark: CoVoST 2

- Reproducing baselines in CoVoST 2 paper
- Many-to-one and one-to-many multilingual models
- Improving low-resource directions with learned features from self-supervised model (wav2vec)

	Fr	De	Es	Zh	Tr	Ar	Sv	Lv	Sl	Ta	Ja	Id	Cy
						X→En							
B-Base	23.2	15.7	20.2	4.4	2.2	2.7	1.4	1.2	1.5	0.2	1.1	1.0	1.7
+ SSL*	23.1	16.2	20.2	4.8	3.2	3.8	3.7	2.3	2.2	0.2	1.6	1.6	2.2
Multi. B-Big†	26.6	19.5	26.3	4.4	2.1	0.3	1.3	0.6	1.4	0.1	0.6	0.3	0.9
T-Sm	26.3	17.1	23.0	5.8	3.6	4.3	2.7	2.5	3.0	0.3	1.5	2.5	2.7
Multi. T-Md†	26.5	17.5	27.0	5.9	2.3	0.4	0.5	0.6	0.7	0.1	0.1	0.3	1.9
						En→X							
B-Base	-	12.5	-	20.0	6.7	9.1	18.1	8.7	11.6	7.4	25.6	15.2	18.9
Multi. B-Big‡	-	12.6	-	22.2	7.3	8.0	18.3	8.9	11.4	7.3	28.2	16.0	19.3
T-Sm	-	16.3	-	25.4	10.0	12.1	21.8	13.0	16.0	10.9	29.6	20.4	23.9
Multi. T-Md‡	-	15.4	-	26.5	9.5	10.8	20.9	12.2	14.6	10.3	30.5	18.9	22.0

Table 5: FAIRSEQ S2T models on CoVoST 2. Test BLEU reported (character-level BLEU for Zh and Ja targets). * Replaced mel-filter bank features with wav2vec ones (Schneider et al., 2019; Wu et al., 2020). † Trained jointly on all 21 X-En directions with temperature-based (T=2) resampling (Arivazhagan et al., 2019a). ‡ Trained jointly on all 15 En-X directions.

Example (LibriSpeech) - Training

```
# Common fairseq-train interface with S2T task and model


fairseq-train ${DATA_ROOT} --task speech_to_text --arch
s2t_transformer_s --criterion label_smoothed_cross_entropy --save-dir
${SAVE_DIR} --max-update 300000 --optimizer adam --lr 2e-3 --lr-
scheduler inverse_sqrt --warmup-updates 10000
```

Example (LibriSpeech) - Decoding

```
# Common fairseq-generate interface with S2T task and scorer


for SUBSET in dev-clean dev-other test-clean test-other; do
  fairseq-generate ${DATA_ROOT} --gen-subset ${SUBSET} --task
    speech_to_text --scoring wer --path ${SAVE_DIR}/checkpoint_best.pt
done
```

Example – S2T Data Prep

- Manifest TSV
  - Audio/feature path, target text, optional data fields
  - One/multiple TSVs per split
- Data configuration YAML (feature transforms, tokenizer, etc.)
- Other data files (e.g. SentencePiece model, global CMVN statistics)
- [Optional] Pre-computed features (.npy)

More Information

- Documentation on Github (fairseq): examples/speech_to_text
- More examples & paper code coming

Thank you! Questions?

FACEBOOK