NAVER CLOVA

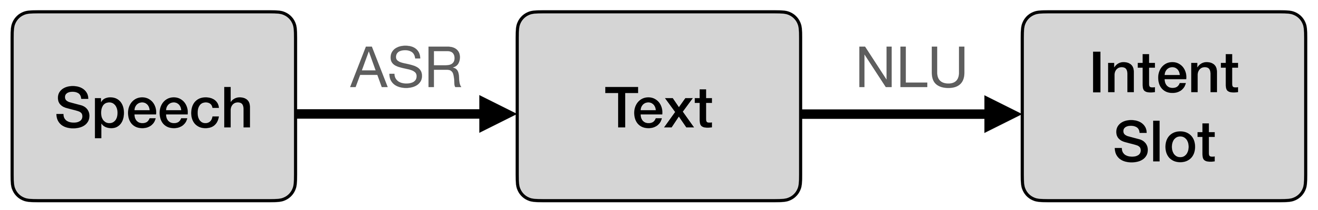
Two-Stage Textual Knowledge Distillation for End-to-End Spoken Language Understanding

Seongbin Kim*, Gyuwan Kim*, Seongjin Shin, Sangmin Lee (*: equal contribution)

#4641

Spoken Language Understanding

- Traditional SLU (ASR + NLU)



(-) prone to error propagation

(+) better interpretability

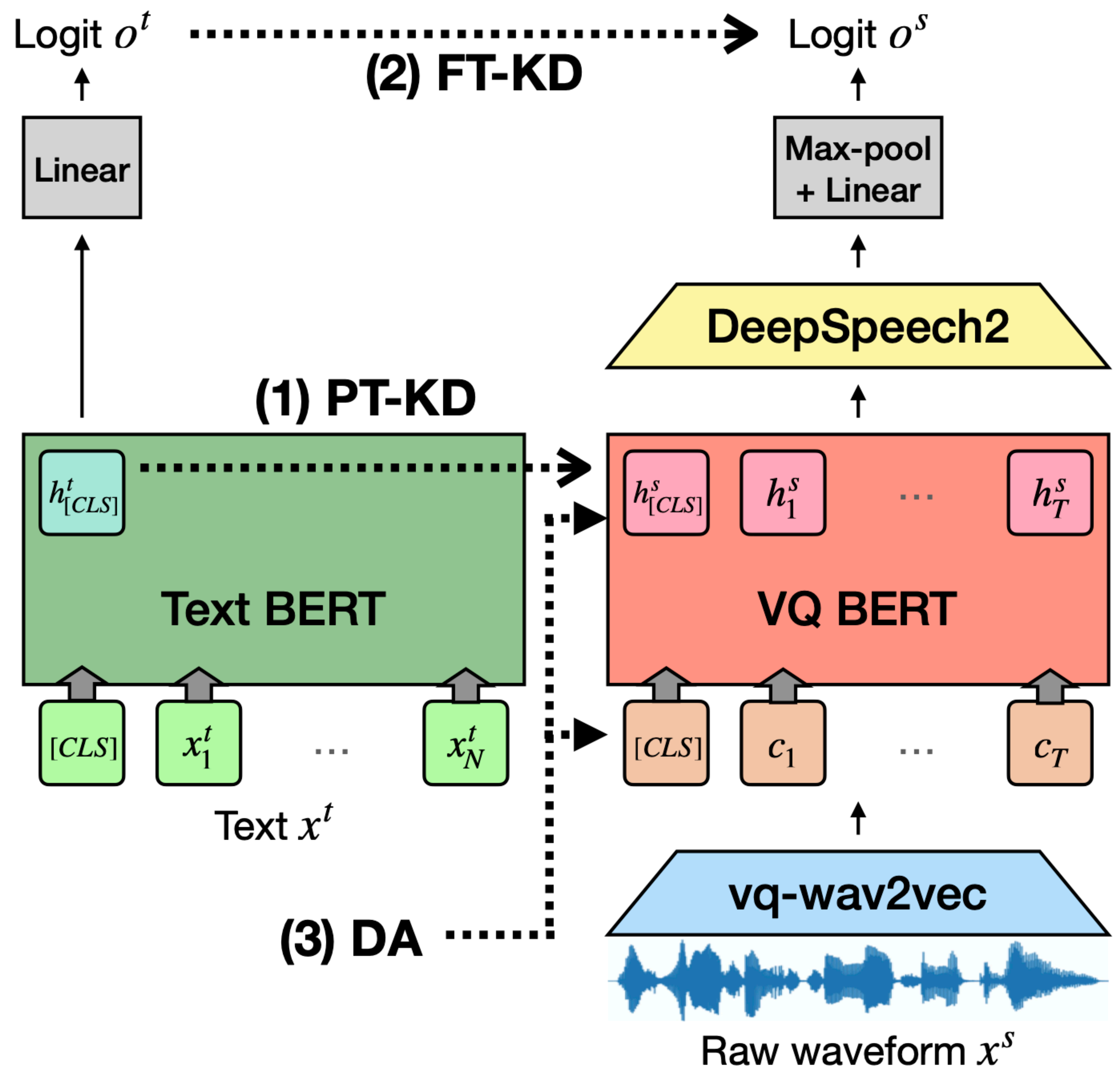
(+) more data/models for ASR/NLU separately

- End-to-end SLU

(-) lack of paired data

(+) more efficient

(+) less information loss (acoustic features)



Model Architecture

- SLU model: vq-wav2vec BERT + DeepSpeech2
- Text model: RoBERTa-base

Training

- Borrow pre-trained vq-wav2vec BERT
  - Fix vq-wav2vec part
- Further pre-training
  - Masked language modeling on 960h of Librispeech
  - Knowledge Distillation (**PT-KD**): sequence-level contextualized representations (L1 loss)
- AM pre-training (**AM-PT**) for better initialization
- Fine-tuning for SLU
  - Intent classification
  - Knowledge Distillation (**FT-KD**): predicted logits (L1 loss)
  - Data augmentation (**DA**): span masking - code, time, channel

Motivation of Our Work

- Small # of SLU training pairs
  - Previous E2E SLU approaches: fine-tuning on SLU after pre-training on ASR
  - Our solution: injecting *textual information* to a speech encoder by *knowledge distillation*
- Recent multi-modal works are successful
  - Most of them are for vision-and-language
  - How about speech-and-language?

Experiments

- Remarkable accuracy in all datasets (especially SOTA performance in FSC Full and 10%)
- All components (PT-KD, FT-KD, AM-PT, DA) are helpful
- DA degrades performance on SNIPS (synthesized data with a single speaker)

	FSC (336)			SNIPS (7)			Smartlights (6)		
	Train	Valid	Test	Train	Valid	Test	Train	Valid	Test
# Speakers	77	10	10	1	1	1	48	2	2
# Utterances	23,132	3,118	3,793	13,084	700	700	1,162	166	332

Method	Full		10%	
	Valid	Test	Valid	Test
Lugosh et al. [6]	-	96.6	-	88.9
+AM-PT [6]	-	98.8	-	97.9
+FT-KD [13]	-	99.0	-	98.1
Price [17]	92.5	99.1	-	-
+DA	94.4	99.4	-	-
+AM-PT	94.8	99.3	-	-
+DA	96.6	99.5	-	-
VQ-BERT +DS2	93.1	98.9	87.3	97.0
+PT-KD	94.1	99.0	90.7	98.5
+FT-KD	96.2	99.6	93.3	99.2
+AM-PT	96.4	99.6	94.3	99.3
+DA	**97.8**	**99.7**	**96.2**	**99.5**

Method	SNIPS	Smartlights	
		Close	Far
VQ-BERT +DS2	86.4	75.9	47.9
+PT-KD	88.3	81.3	51.2
+FT-KD	95.3	84.6	59.6
+AM-PT	**96.7**	92.2	70.5
+DA	95.7	**95.5**	**75.0**

Takeaway: (1) using **textual information** for training SLU model is helpful
(2) how to train with texts matters (in our case, **knowledge distillation**)