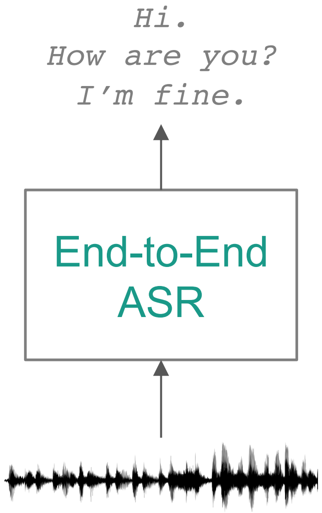
End-to-End Processing of Conversations

Suyoun Kim, Thesis Defense, November 26, 2019

Thesis committees:
Florian Metze (Chair, Carnegie Mellon University)
Richard M. Stern (Co-Chair, Carnegie Mellon University)
Bhiksha Raj (Carnegie Mellon University)
Michael L. Seltzer (Facebook)
Shinji Watanabe (Johns Hopkins University)

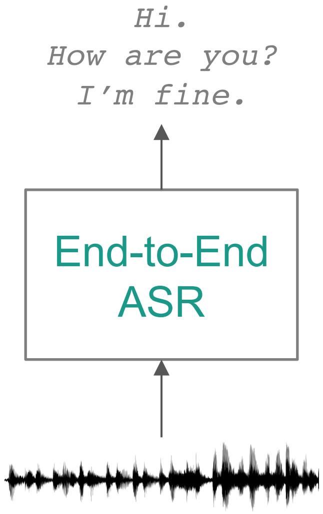
My research projects



My research projects

E2E ASR with Attention for Multi-CH

Kim et al, ICLR workshop 2016; Kim et al, INTERSPEECH 2016



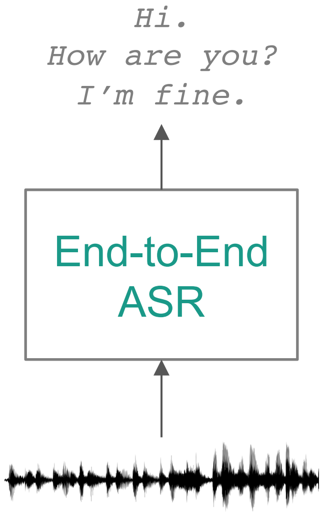
My research projects

E2E ASR with Attention for Multi-CH

Kim et al, ICLR workshop 2016; Kim et al, INTERSPEECH 2016

E2E ASR with joint CTC/Seq2Seq

Kim et al, ICASSP 2017; Watanabe et al, IEEE JSTSP 2017



My research projects

E2E ASR with Attention for Multi-CH

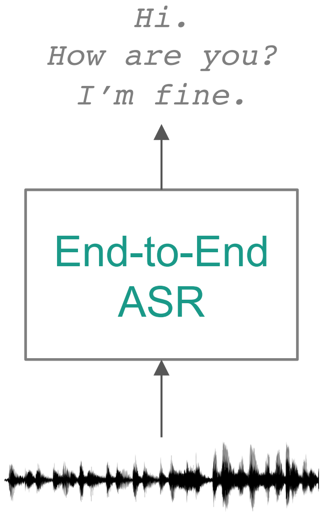
Kim et al, ICLR workshop 2016; Kim et al, INTERSPEECH 2016

E2E ASR with joint CTC/Seq2Seq

Kim et al, ICASSP 2017; Watanabe et al, IEEE JSTSP 2017

E2E ASR for Multi-Language & online ASR

Kim et al, ICASSP 2018; Kim et al, INTERSPEECH 2018



My research projects

E2E ASR with Attention for Multi-CH

Kim et al, ICLR workshop 2016; Kim et al, INTERSPEECH 2016

E2E ASR with joint CTC/Seq2Seq

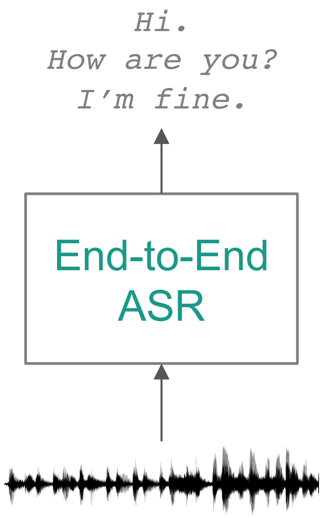
Kim et al, ICASSP 2017; Watanabe et al, IEEE JSTSP 2017

E2E ASR for Multi-Language & **online ASR**

Kim et al, ICASSP 2018; Kim et al, INTERSPEECH 2018

E2E ASR for Conversations

Kim et al, CHiME5 2018; Kim et al, SLT 2018; Kim et al, NAACL 2019; Kim et al, ACL 2019; Kim et al, INTERSPEECH 2019



Today's Talk is ...

E2E ASR with Attention for Multi-CH

Kim et al, ICLR workshop 2016; Kim et al, INTERSPEECH 2016

E2E ASR with joint CTC/Seq2Seq

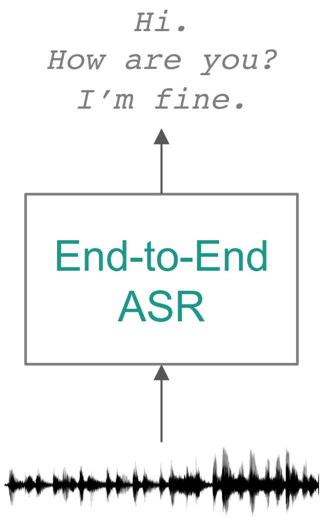
Kim et al, ICASSP 2017; Watanabe et al, IEEE JSTSP 2017

E2E ASR for Multi-Language & **online ASR**

Kim et al, ICASSP 2018; Kim et al, INTERSPEECH 2018

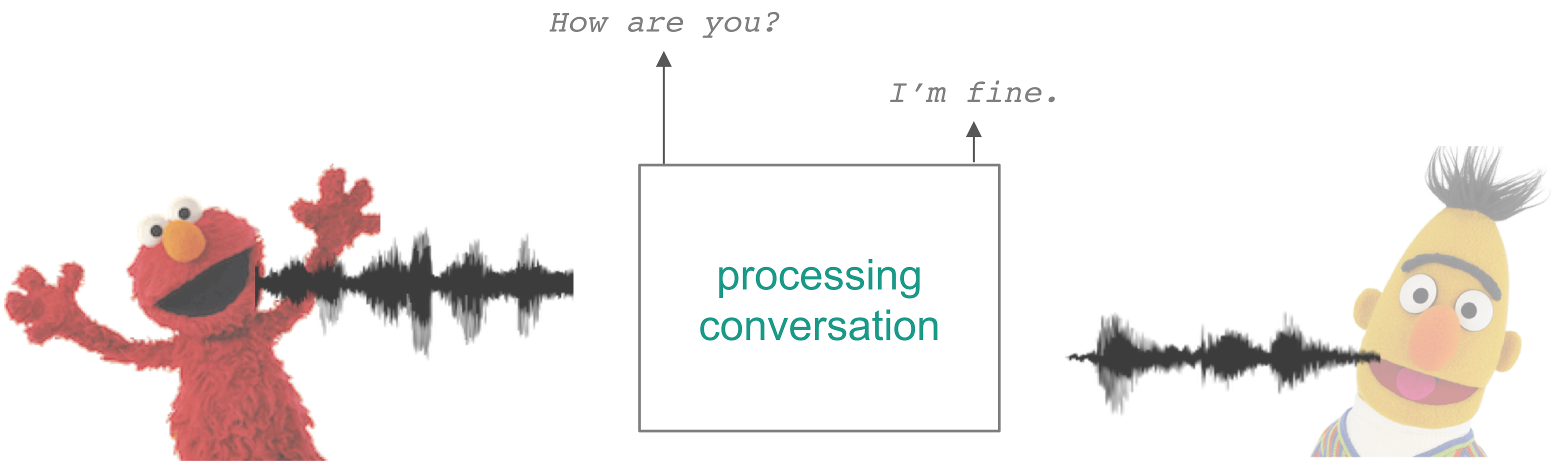
E2E ASR for Conversations

Kim et al, CHiME5 2018; Kim et al, SLT 2018; Kim et al, NAACL 2019; Kim et al, ACL 2019; Kim et al, INTERSPEECH 2019



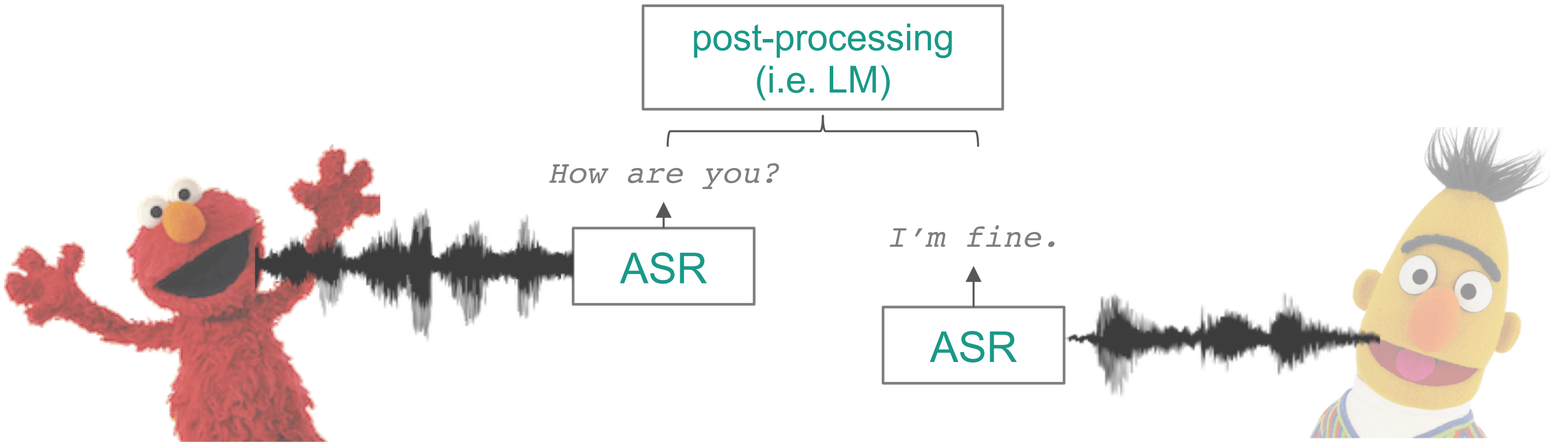
Processing of conversations is a core technique in "Conversational AI"

Analysis of meeting, conversations, interviews, etc…



Current ASR solutions, even state-of-the-art systems, are modeling fragments, not conversations

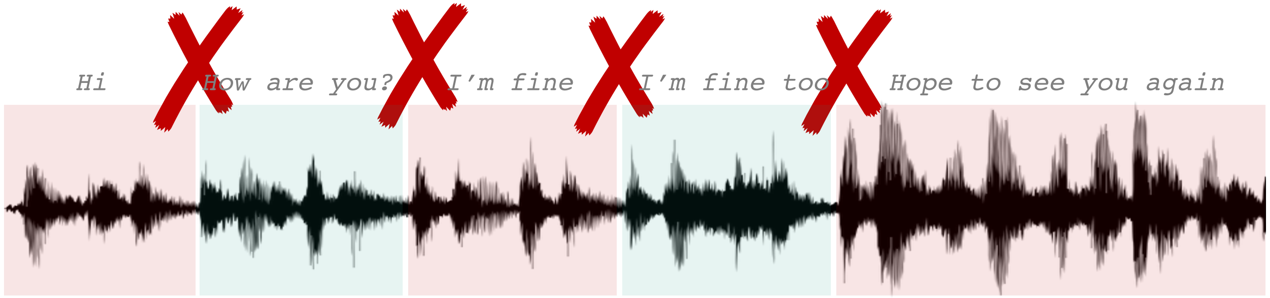
Conversation is split into utterances, then ASR is built on that utterances



Current ASR solutions lose long context, beyond utterances

ASR cannot learn dependencies between utterances

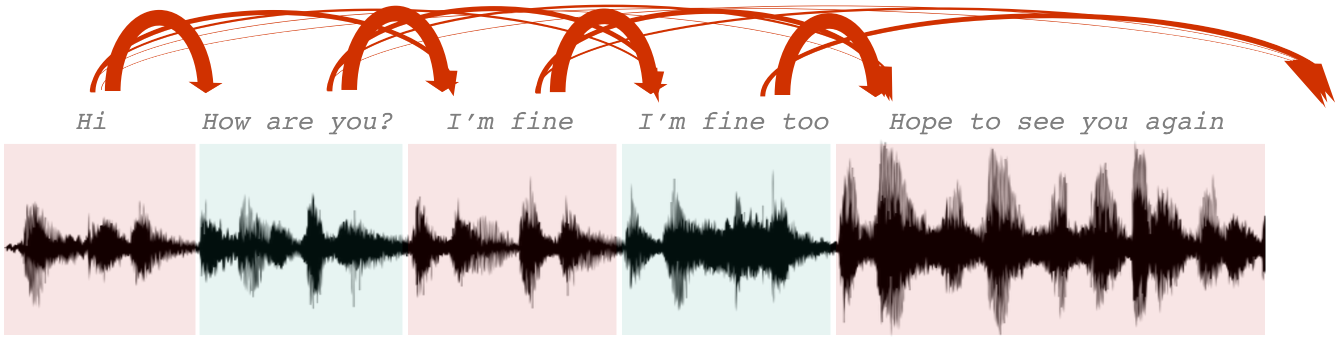
(Long context is only added as postfix by LM)



My goal is to use "conversational context" in End-to-End ASR

"Conversational context" = higher-level knowledge that spans across multiple utterances, which is helpful to process long conversation

- Similar words/ phrases tend to repeat
- Topical coherence tend to exist



Previous studies for using conversational context has been only explored in language modeling

- Dialog session-based LM
  - by Xiong et al. (MS) in 2017
- Turn-based Dialog context LM
  - by Liu et al. (CMU) in 2017

➔ Conversational context knowledge still added as postfix

- Contextual End-to-End ASR
  - by Pundak et al. (Google) in 2018

➔ This context is about user-specific phrases (e.g. contact lists, song lists), not a long, conversational context

Bringing in conversational context into ASR

I propose,

1. Efficient way to ***preserve*** long conversational context while overcoming GPU memory issue
   (Kim et al, SLT 2018; Kim et al, NAACL 2019)
2. Effective way to ***integrate*** conversational context into ASR
   (Kim et al, ACL 2019)
3. Methods to ***encode*** conversational context
   (Kim et al, ACL 2019; Kim et al, INTERSPEECH 2019)
   - using previous spoken utterances
   - using external "world knowledge" of word/sentence

Overview

- ❑ How to preserve and integrate "conversational context"?
- ❑ How to encode "conversational context"?
- ❑ Experiments and Analysis

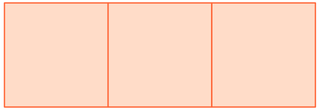
Overview

- ❑ **How to preserve and integrate "conversational context"?**
- ❑ **How to encode "conversational context"?**
- ❑ **Experiments and Analysis**

Simplest way could be treating an entire conversation as an utterance to preserve conversational context…

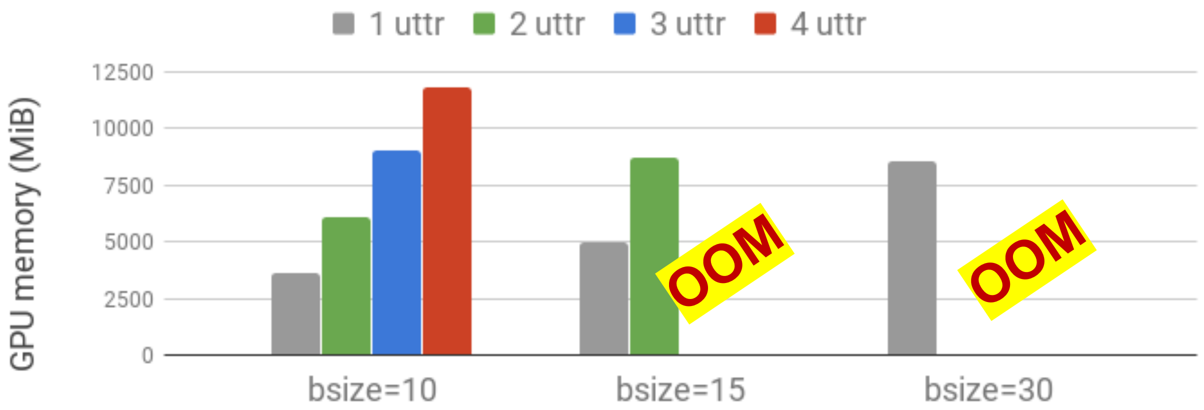
However,

- Speech input feature sequences are already too long
  (1 sec has ~30 frames and ~10 char, then 20 mins dialogs?)
- Simply concatenating utterances (like LM) won't work
  Slow to train, cannot fit in GPU memory
  Poor parallelization due to severely variable-length of each dialog

➔ We need to **`extract`** some sort of embeddings as "context" : 

BPTT on entire conversation is computationally infeasible

We conducted a simple experiment



TITAN X (Pascal) ~11G
300h-SWBD with batch-size=20
takes 3 hrs/ epoch and 20 epochs

➔ We need to **detach** the graph for context until needed (like truncated BPTT)

We extract & detach & cache contexts on serialized minibatches

We create minibatches with serialized utterances based on their start time and apply randomization only at dialog level

Kim et al, SLT 2018

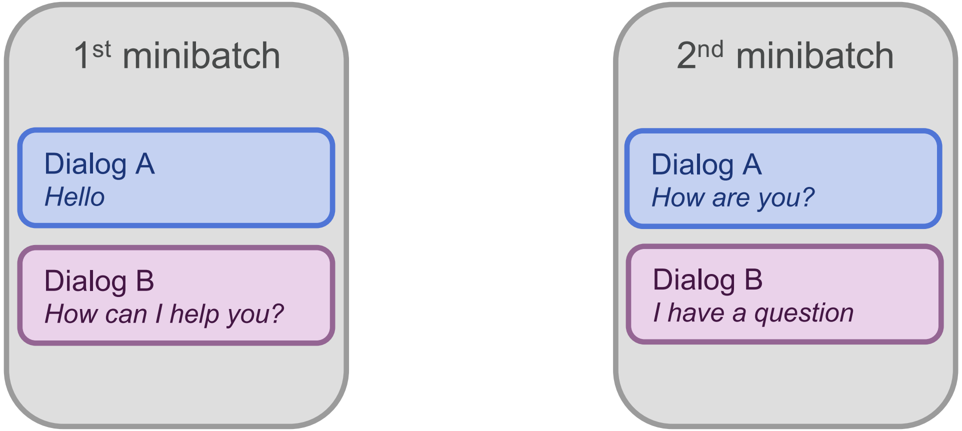
We extract & detach & cache contexts on serialized minibatches

We create minibatches with serialized utterances based on their start time and apply randomization only at dialog level

Kim et al, SLT 2018

We extract & detach & cache contexts on serialized minibatches

We create minibatches with serialized utterances based on their start time and apply randomization only at dialog level



Kim et al, SLT 2018

We extract & detach & cache contexts on serialized minibatches

We create minibatches with serialized utterances based on their start time and apply randomization only at dialog level

Kim et al, SLT 2018

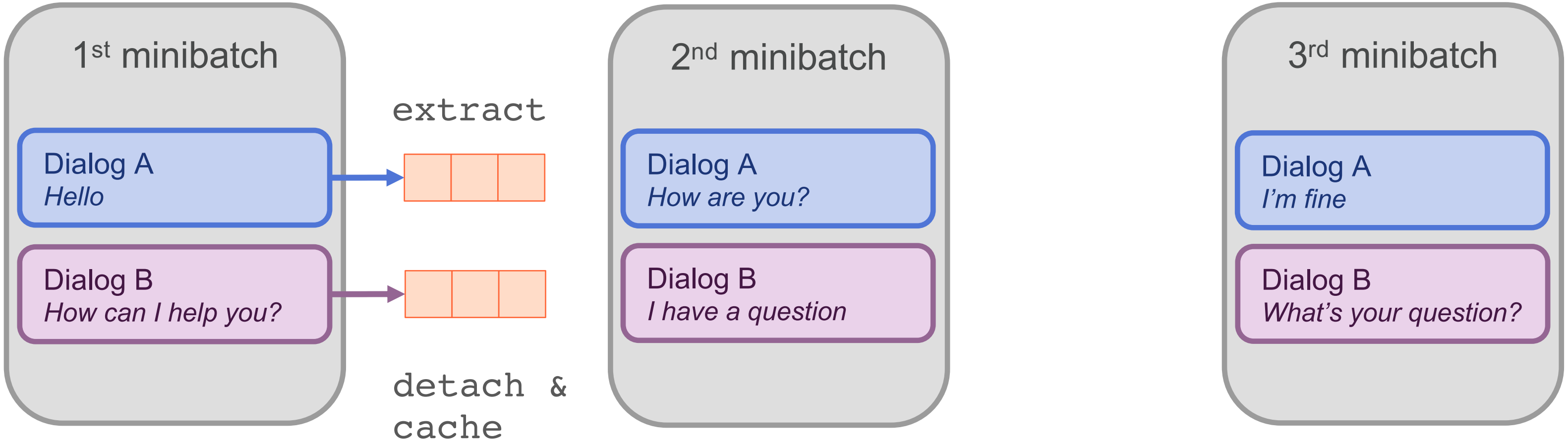
We extract & detach & cache contexts on serialized minibatches

We create minibatches with serialized utterances based on their start time and apply randomization only at dialog level

Kim et al, SLT 2018

We extract & detach & cache contexts on serialized minibatches

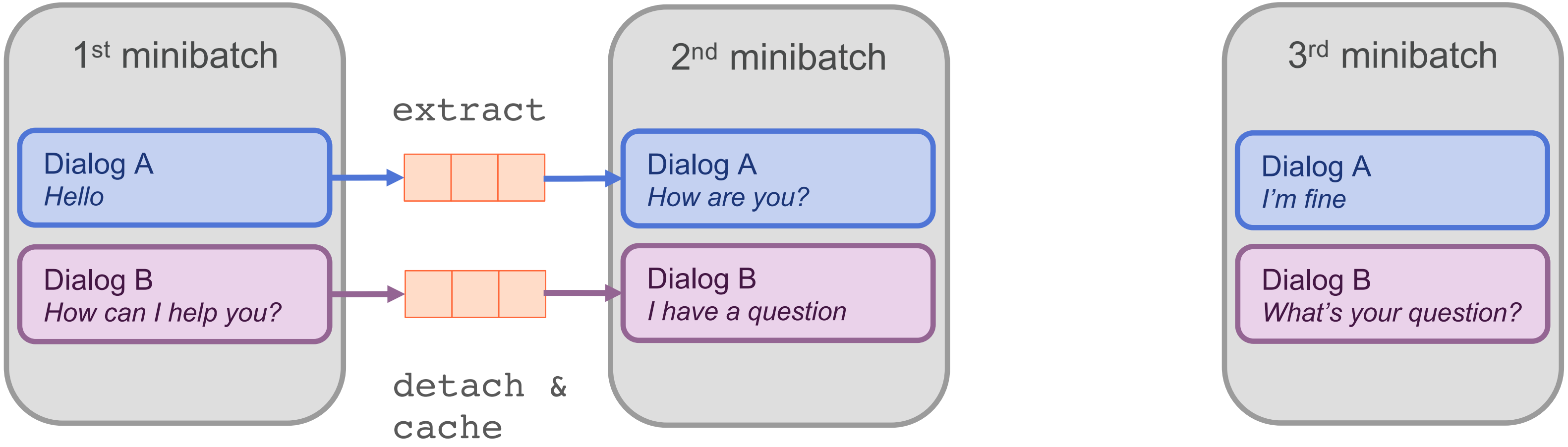
We create minibatches with serialized utterances based on their start time and apply randomization only at dialog level



Kim et al, SLT 2018

We extract & detach & cache contexts on serialized minibatches

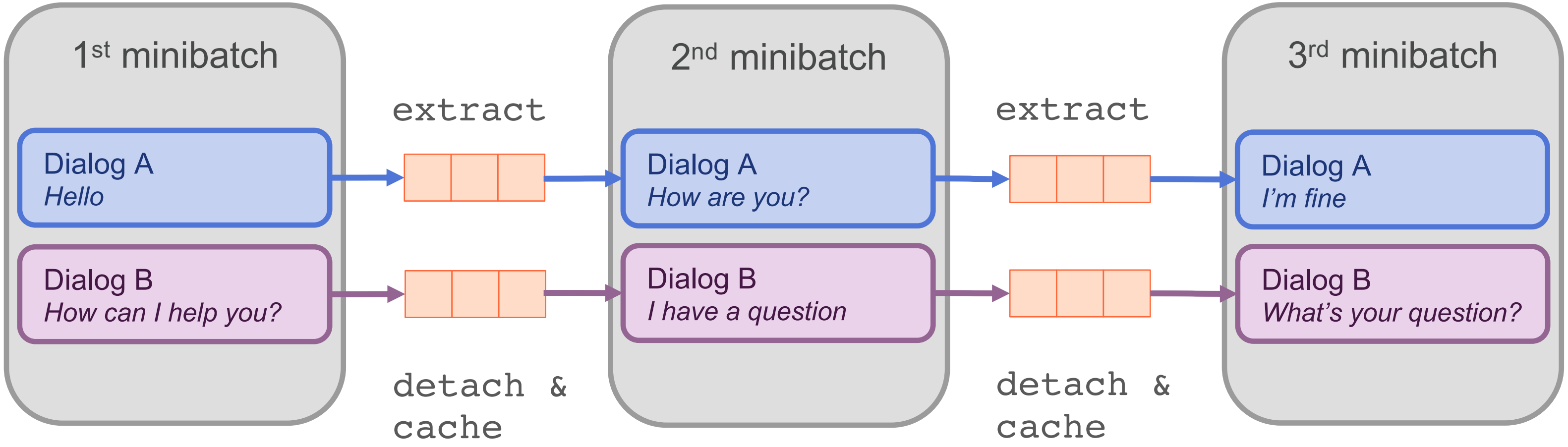
We create minibatches with serialized utterances based on their start time and apply randomization only at dialog level



Kim et al, SLT 2018

We extract & detach & cache contexts on serialized minibatches

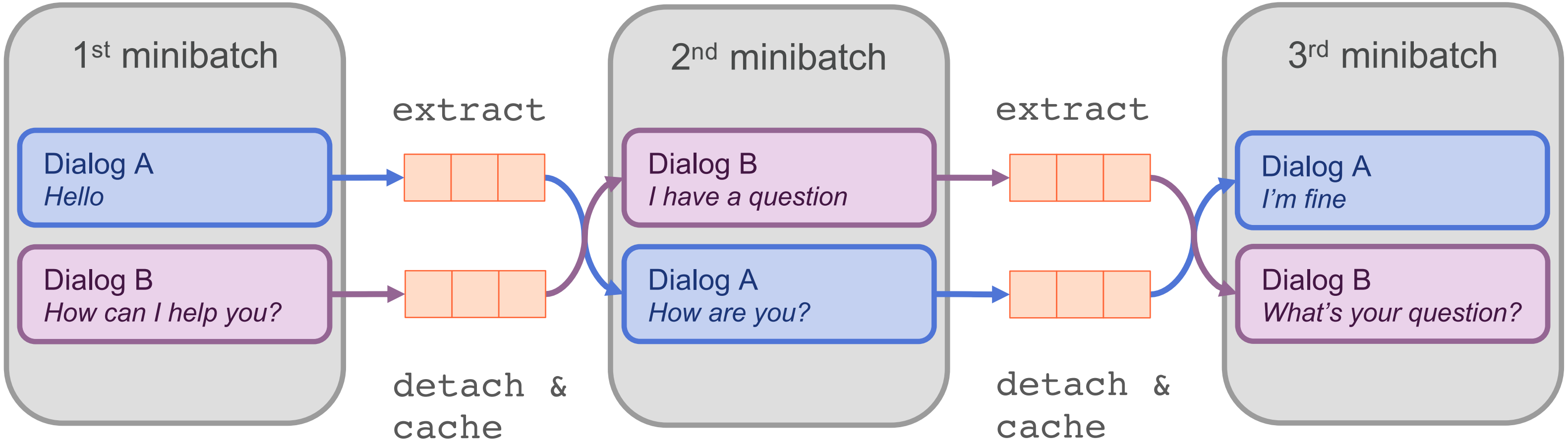
We create minibatches with serialized utterances based on their start time and apply randomization only at dialog level



Kim et al, SLT 2018

We extract & detach & cache contexts on serialized minibatches

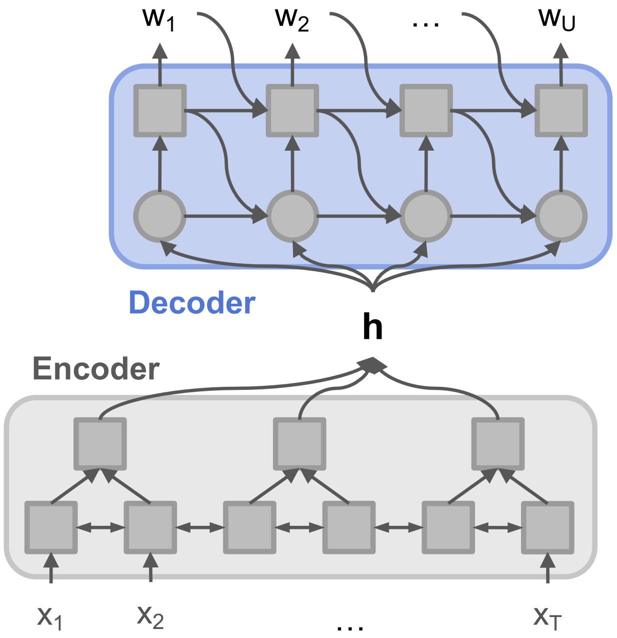
We create minibatches with serialized utterances based on their start time and apply randomization only at dialog level



Kim et al, SLT 2018

End-to-End ASR model that we use as basis

- $x_{1:T}$ = input
- $w_{1:U}$ = output
- **h** = high-level speech feature



Typical End-to-End Model (Seq2Seq, LAS)

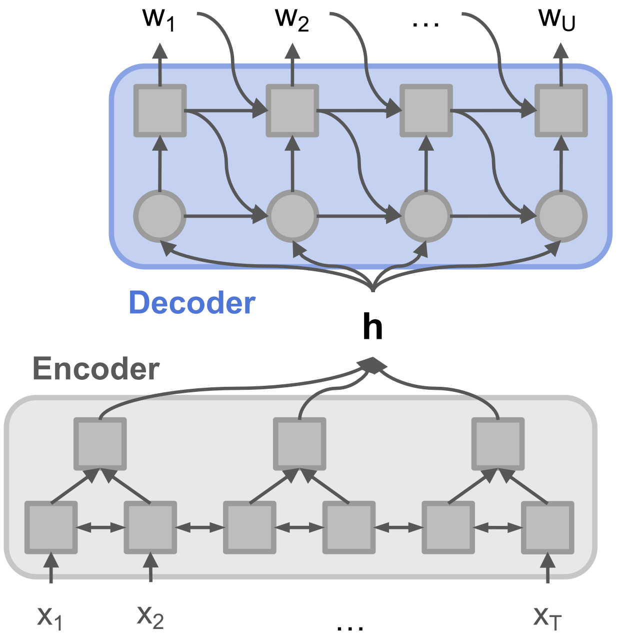
Chan et al, ICASSP 2016
Bahdanau et al, ICASSP 2016

End-to-End ASR model that we use as basis

- $x_{1:T}$ = input
- $w_{1:U}$ = output
- **h** = high-level speech feature

- Learning

$$\max_{\theta} P(w_{1:U} | x_{1:T}; \theta)$$

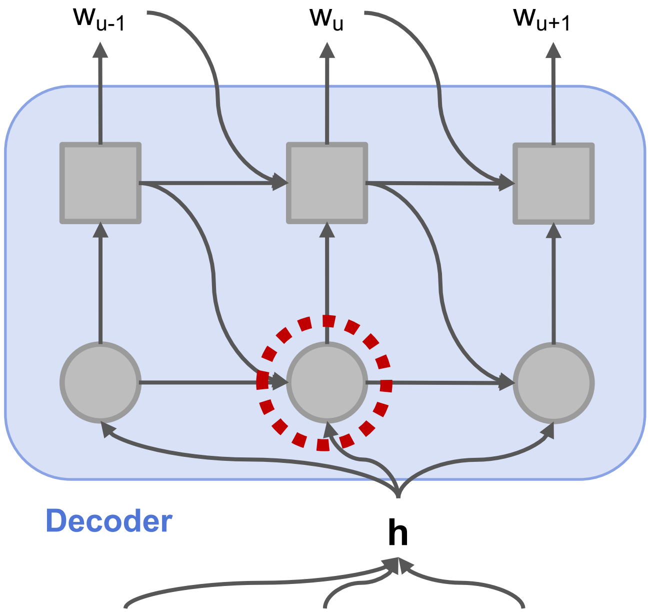


Typical End-to-End Model (Seq2Seq, LAS)

Chan et al, ICASSP 2016
Bahdanau et al, ICASSP 2016

Decoder part of End-to-End ASR with attention mechanism

- **h** = high-level speech feature

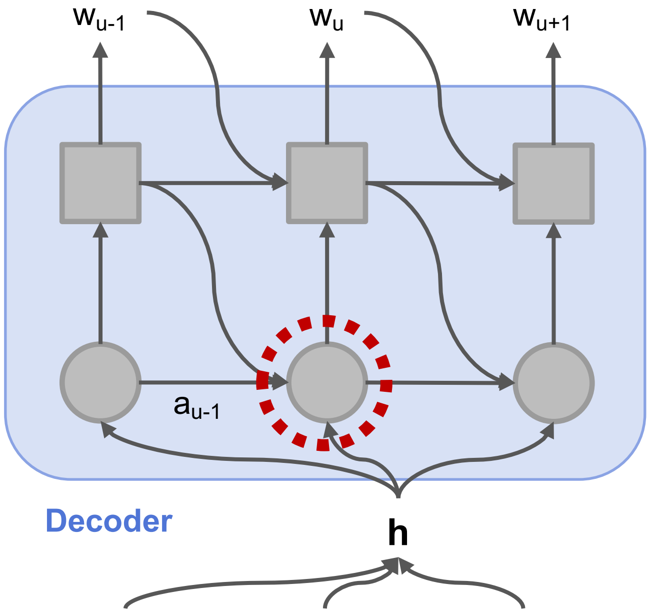


Typical End-to-End Model – Decoder part

Chan et al, ICASSP 2016
Bahdanau et al, ICASSP 2016

Decoder part of End-to-End ASR with attention mechanism

- **h** = high-level speech feature
- a_{u-1} = previous attention

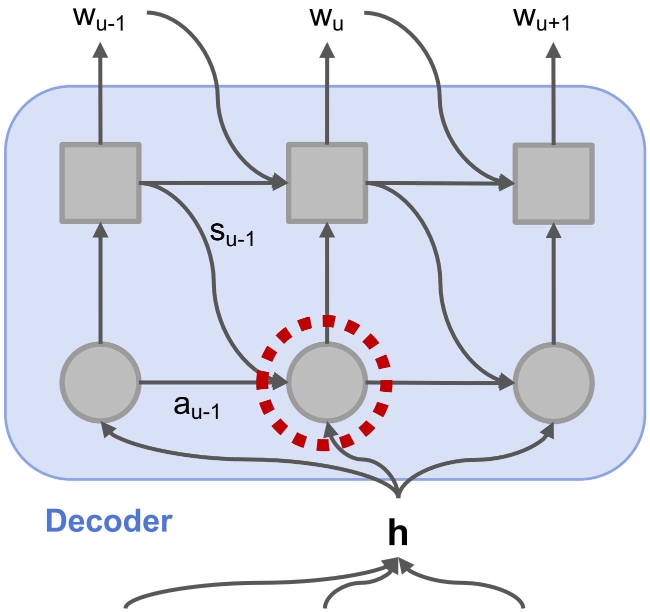


Typical End-to-End Model – Decoder part

Chan et al, ICASSP 2016
Bahdanau et al, ICASSP 2016

Decoder part of End-to-End ASR with attention mechanism

- **h** = high-level speech feature
- a_{u-1} = previous attention
- s_{u-1} = previous decoder state

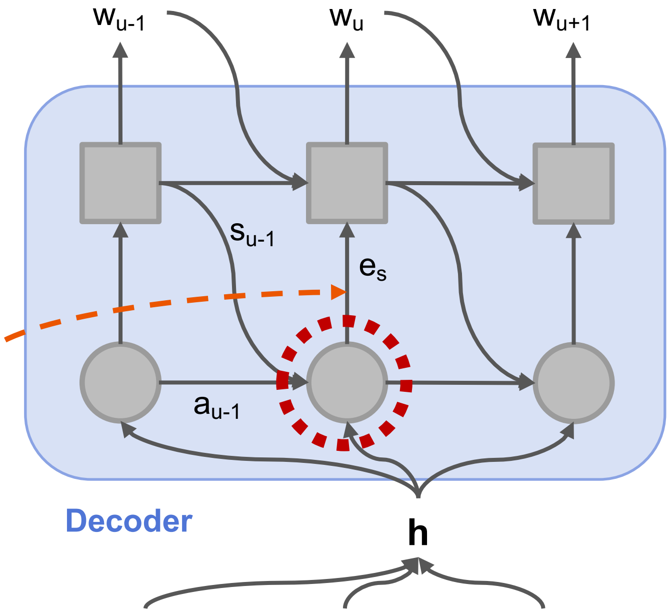


Typical End-to-End Model – Decoder part

Chan et al, ICASSP 2016
Bahdanau et al, ICASSP 2016

Decoder part of End-to-End ASR with attention mechanism

- **h** = high-level speech feature
- a_{u-1} = previous attention
- s_{u-1} = previous decoder state
- Attention mechanism already has "context" vector, let's call it as speech embedding, e_s

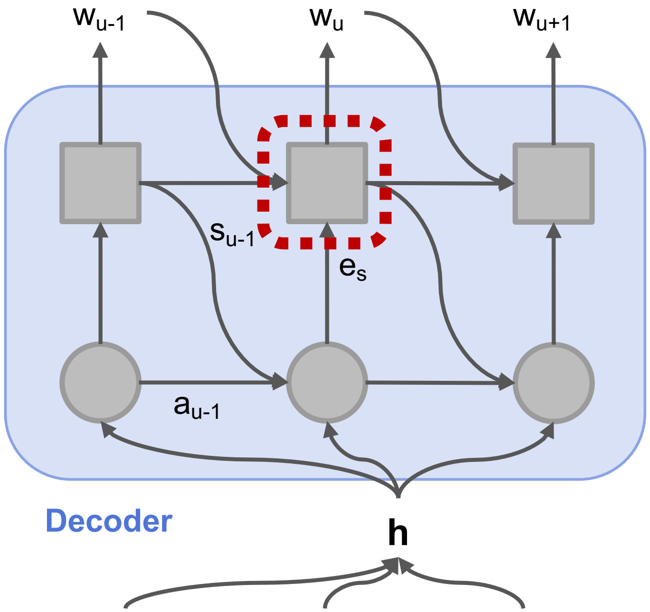


Typical End-to-End Model – Decoder part

Chan et al, ICASSP 2016
Bahdanau et al, ICASSP 2016

Decoder part of End-to-End ASR takes two different types of embeddings: word, speech

- e_s = speech embedding

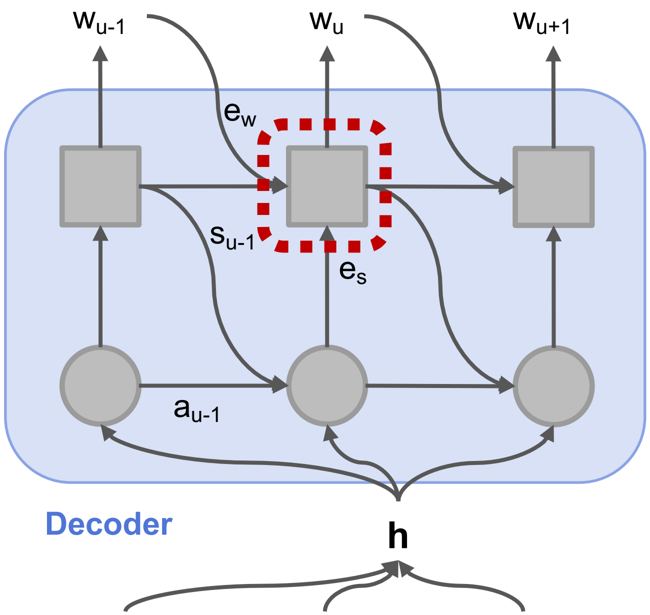


Typical End-to-End Model – Decoder part

Chan et al, ICASSP 2016
Bahdanau et al, ICASSP 2016

Decoder part of End-to-End ASR takes two different types of embeddings: word, speech

- e_s = speech embedding
- e_w = word embedding



Typical End-to-End Model – Decoder part

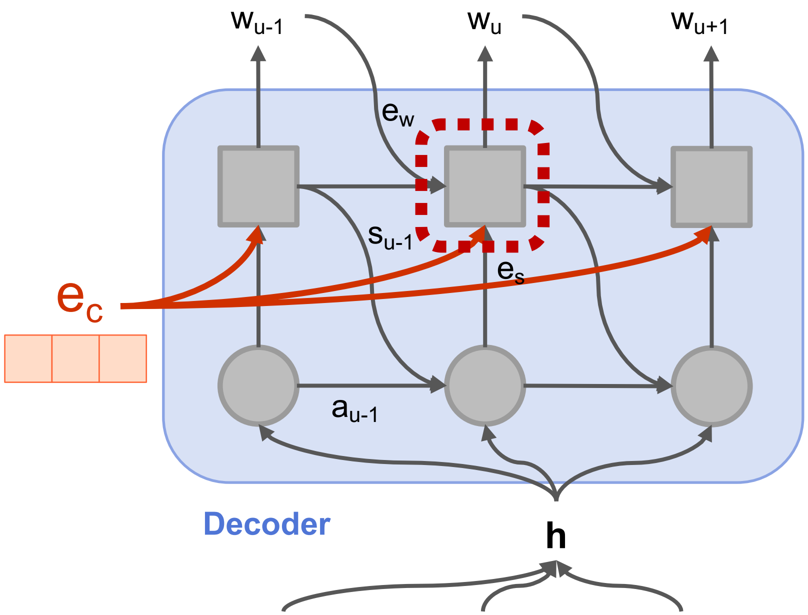
Chan et al, ICASSP 2016
Bahdanau et al, ICASSP 2016

We extend decoder part of End-to-End ASR since we now have "context embedding"

- e_s = speech embedding
- e_w = word embedding

AND

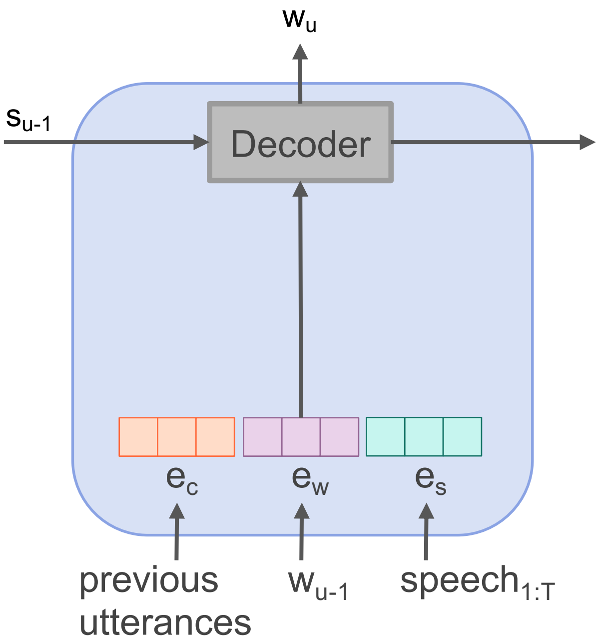
- e_c = context embedding from previous spoken utterances



Kim et al, SLT 2018
Kim et al, NAACL 2019

We propose to use gate mechanism to integrate different types of embeddings: context, word, speech

Rather than simply concatenating

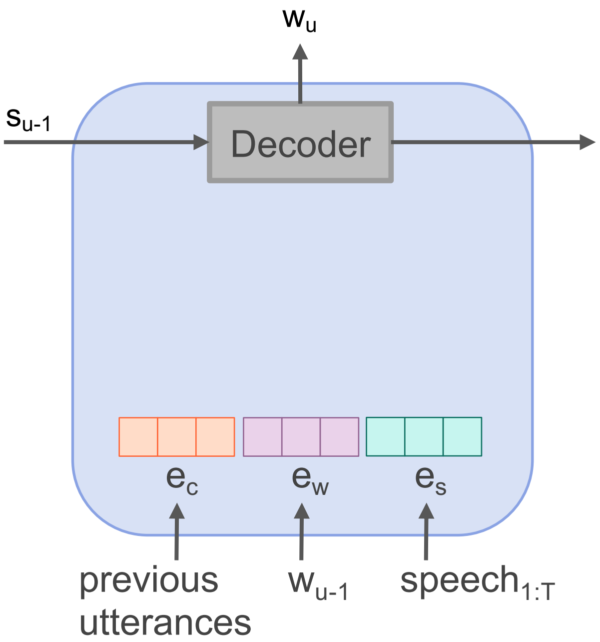


Kim et al, ACL 2019

We propose to use gate mechanism to integrate different types of embeddings: context, word, speech

Rather than simply concatenating

- ➢ Gating mechanism decides how to weigh different embeddings
- ➢ Shape information flow using multiplicative interactions



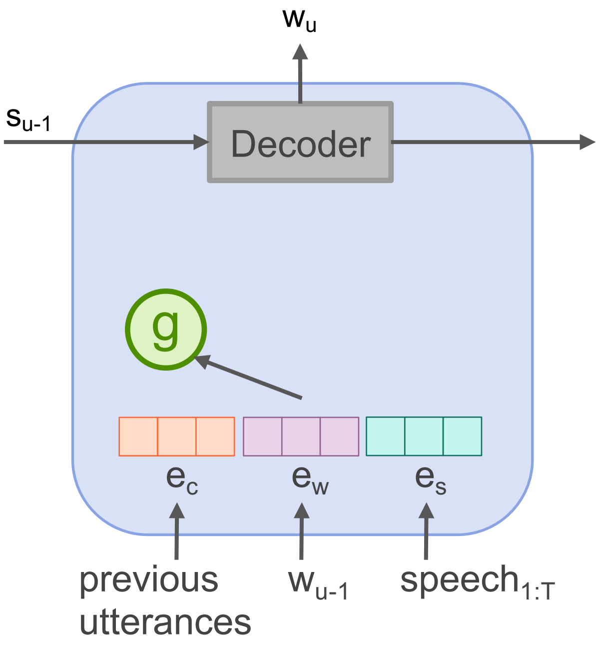
Kim et al, ACL 2019

We propose to use gate mechanism to integrate different types of embeddings: context, word, speech

Rather than simply concatenating

- ➢ Gating mechanism decides how to weigh different embeddings
- ➢ Shape information flow using multiplicative interactions

$$g = \sigma(e_c, e_w, e_s)$$



Kim et al, ACL 2019

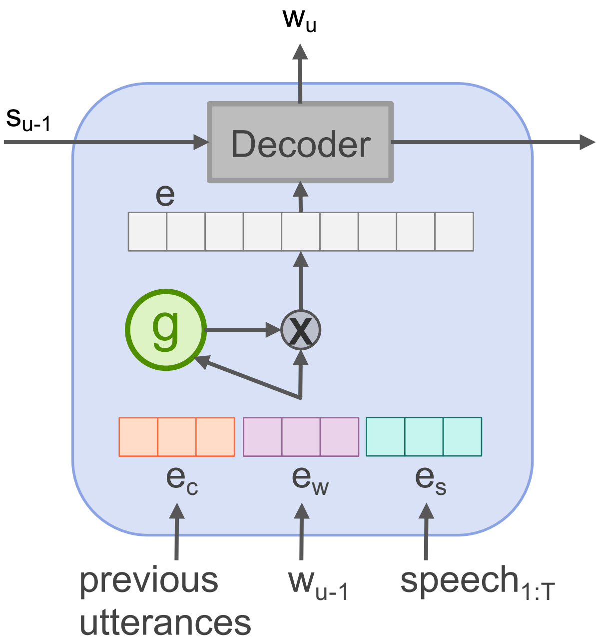
We propose to use gate mechanism to integrate different types of embeddings: context, word, speech

Rather than simply concatenating

- Gating mechanism decides how to weigh different embeddings
- Shape information flow using multiplicative interactions

$$g = \sigma(e_c, e_w, e_s)$$

$$e = g \odot (e_c, e_w, e_s)$$

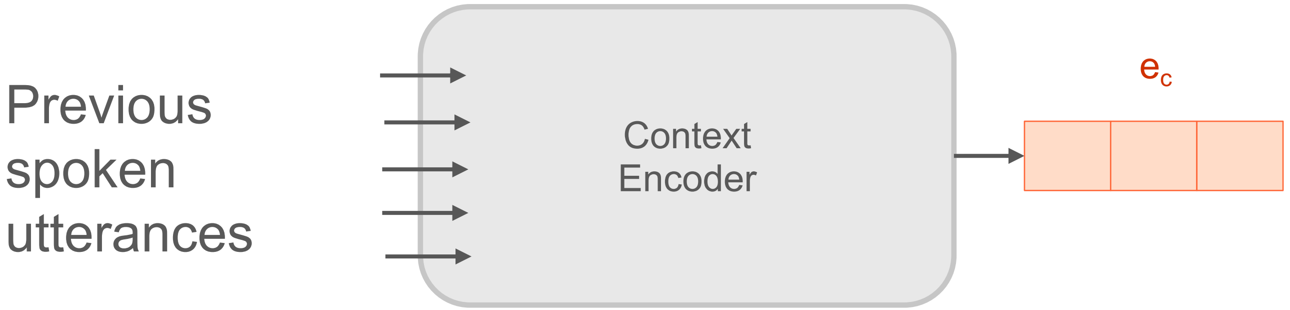


Kim et al, ACL 2019

Overview

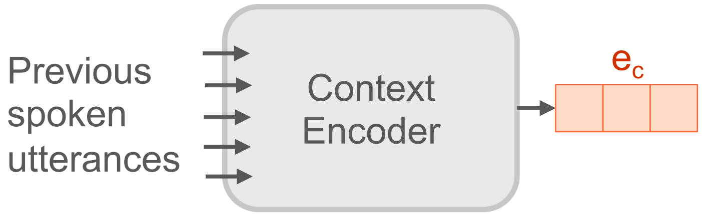
- ❑ How to preserve and integrate "conversational context"?
- ❑ How to encode "conversational context"?
- ❑ Experiments and Analysis

We create "Context Encoder" to map previous spoken utterances into context embedding



Kim et al, SLT 2018
Kim et al, NAACL 2019

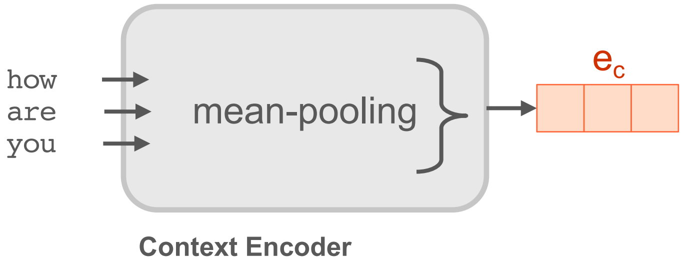
We propose various types of context encoder



1. History units:
   - utterance vs. word
2. Unit representation:
   - utterance vs. word embeddings
   - external embedding
3. Aggregation of history:
   - mean-pooling (simple) vs. RNN vs. Attention function
4. Sampling:
   - ground-truth vs. model output

Kim et al, ACL 2019
Kim et al, Interspeech 2019

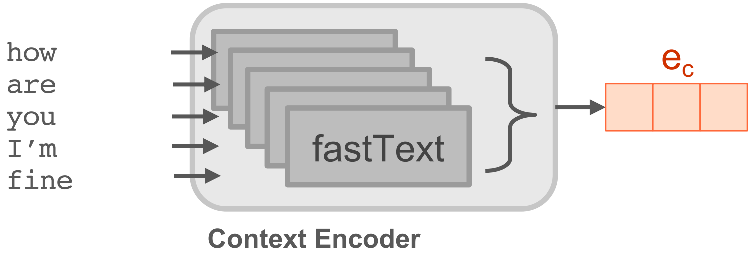
Our method 1) “vanilla” conv.E2E uses word-level unit, mean-pooling, single utterance history



- “vanilla” conv.E2E
  + with gate mechanism
- “vanilla” conv.E2E
  + without gate mechanism

Kim et al, ACL 2019
Kim et al, Interspeech 2019

Our method 2) “fastText / BERT for context” use external “World Knowledge”, multiple histories

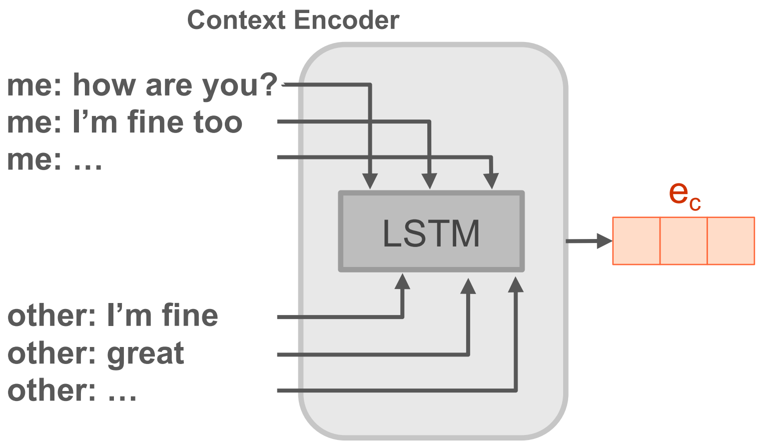


- **fastText for context:** each preceding word is mapped to 300d vector
  + Gate mechanism

- **BERT for context:** each preceding utterance is mapped to 786d vector
  + Gate mechanism

Kim et al, ACL 2019
Kim et al, Interspeech 2019

Our method 3) “LSTM-Attention” for 2-Party Conversations

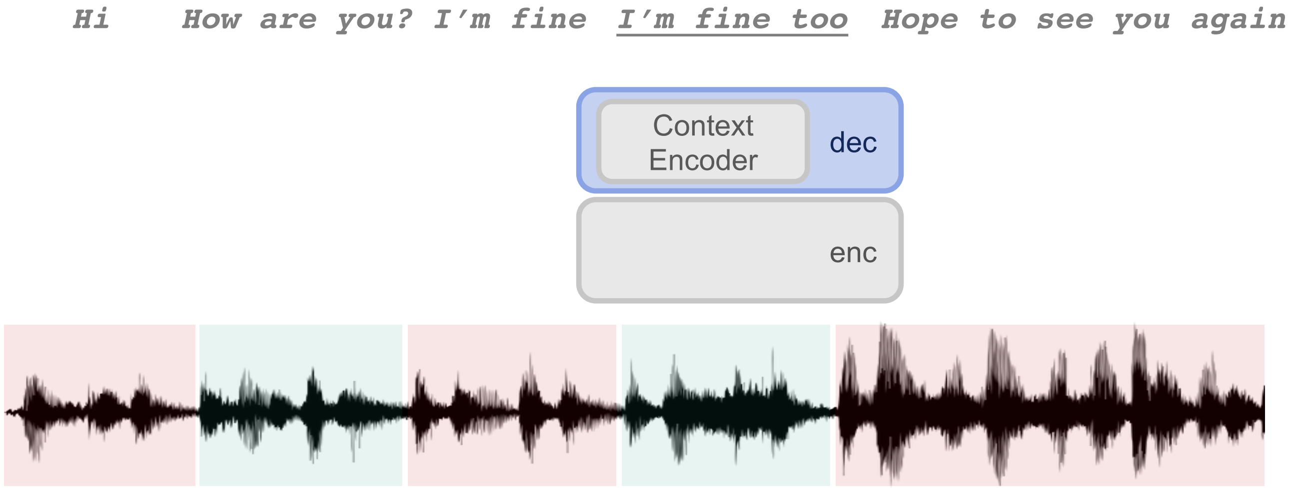


- Consider the turn-change information or interaction between two-speakers

- Learn from history ***what other speaker said*** and ***what current speaker said***.

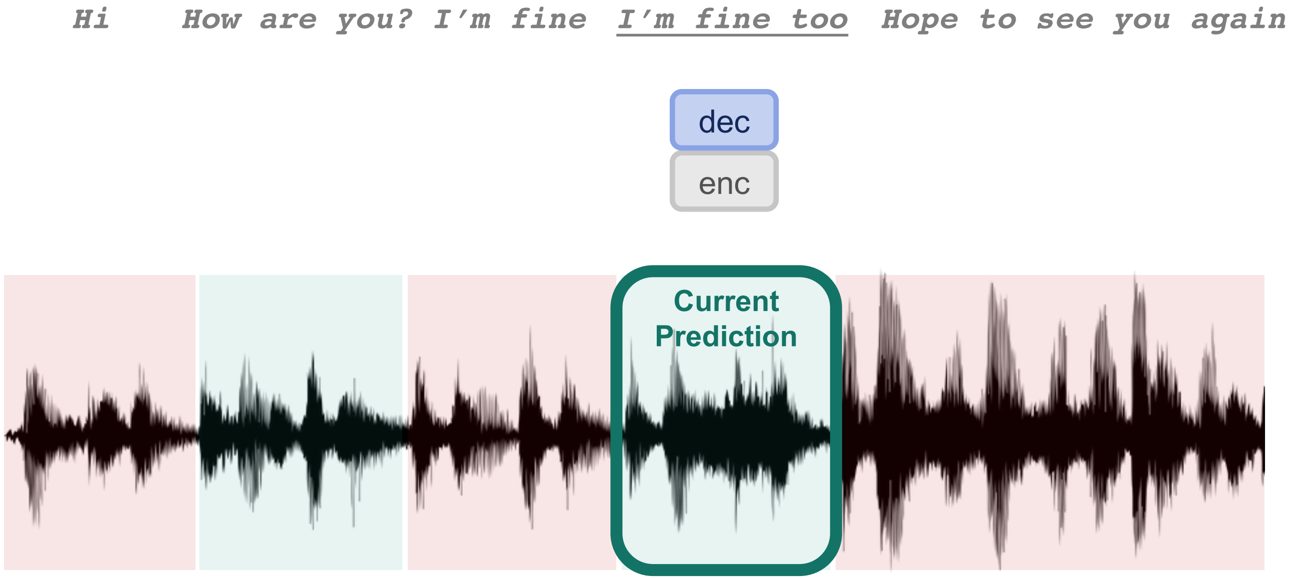
Kim et al, Interspeech 2019

Our context encoder is designed to be trained over all (or window) of past utterances

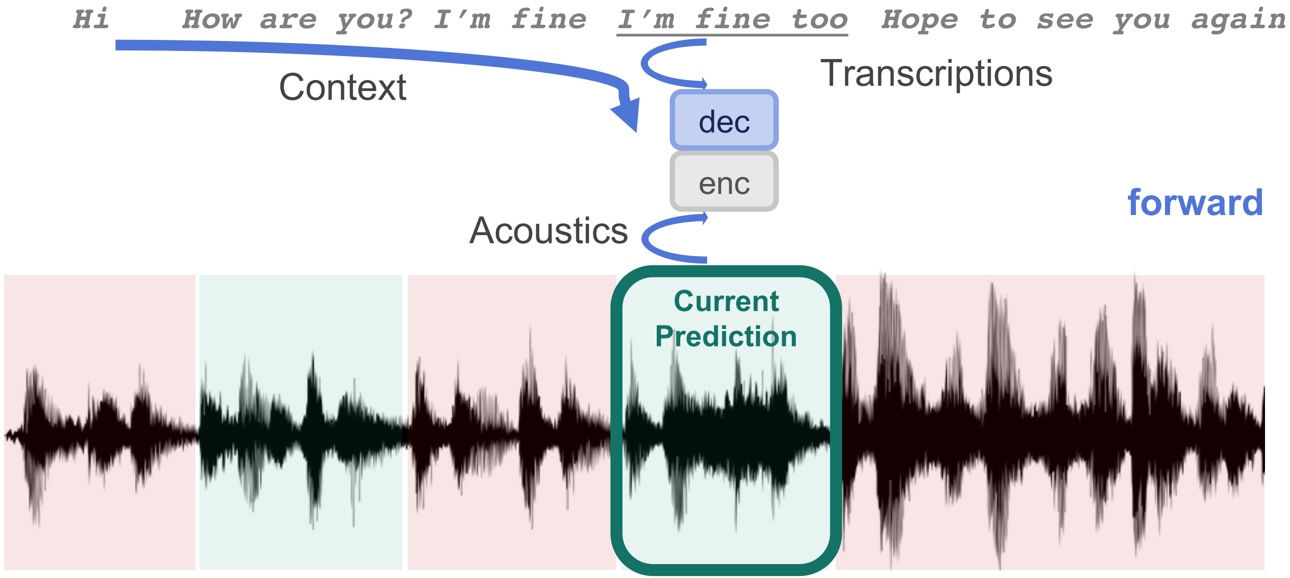


Kim et al, SLT 2018
Kim et al, NAACL 2019
Kim et al, ACL 2019
Kim et al, Interspeech 2019

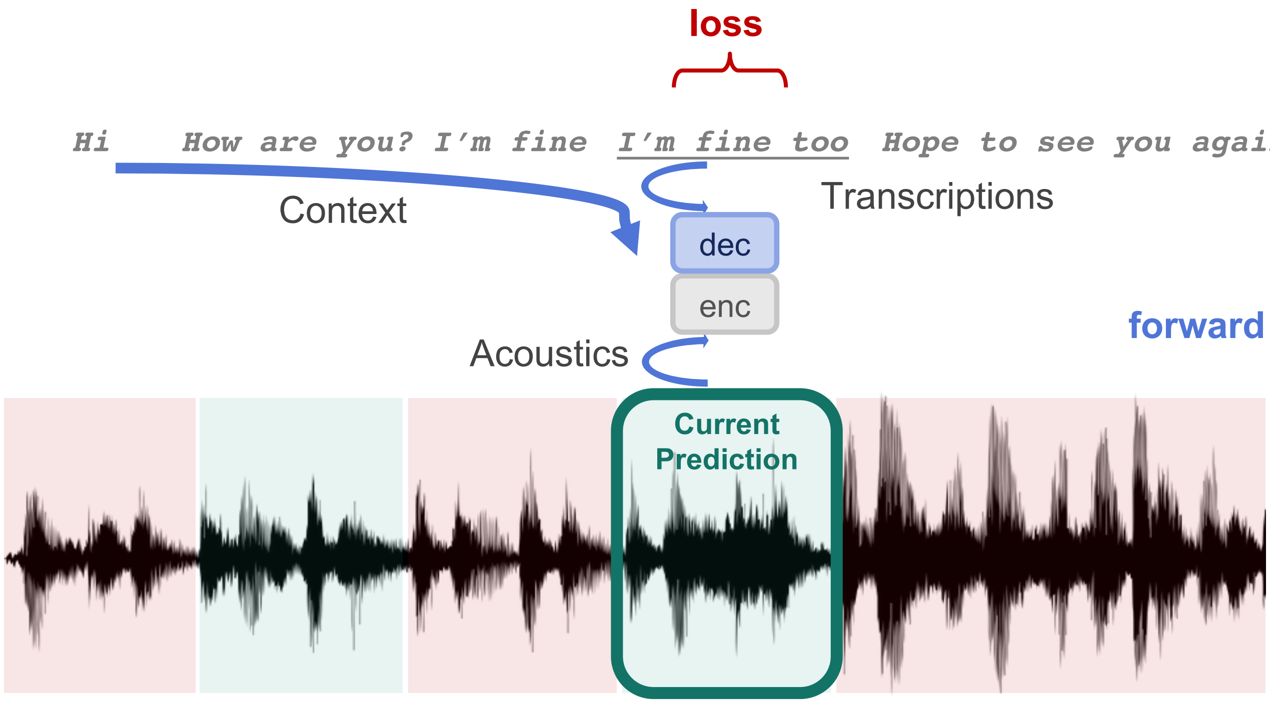
Our context encoder is designed to be trained over all (or window) of past utterances



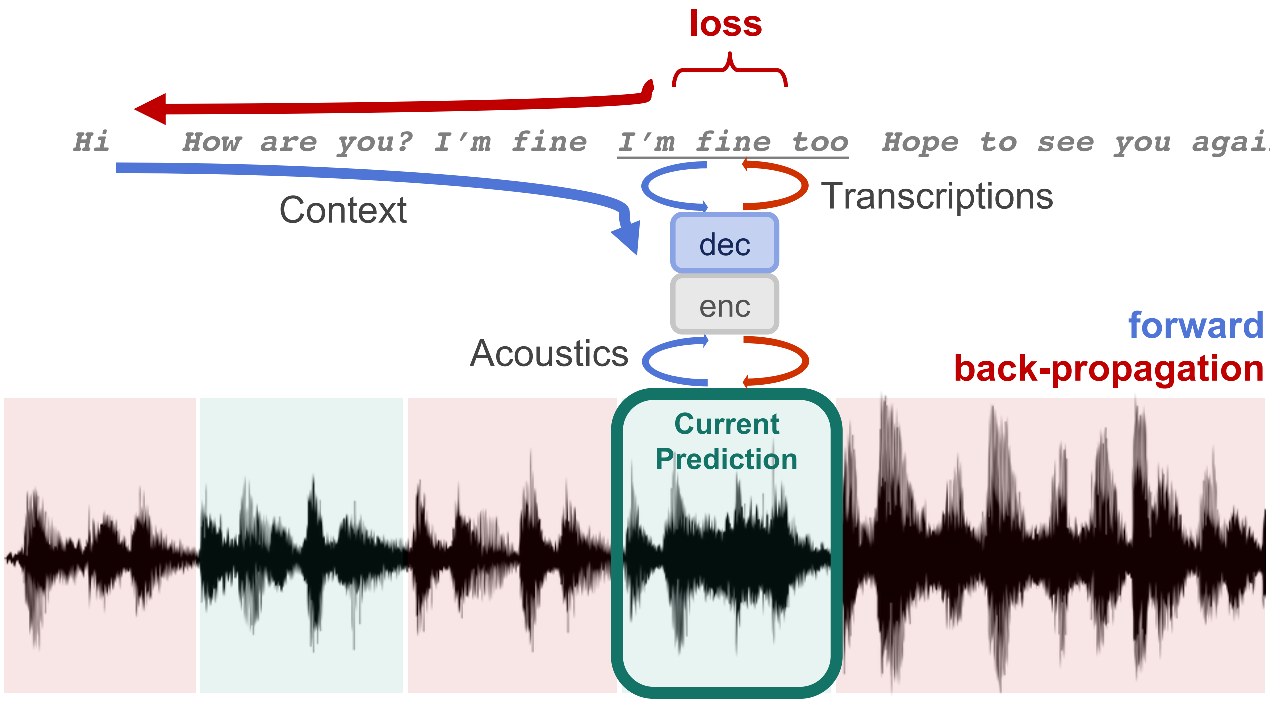
Our context encoder is designed to be trained over all (or window) of past utterances



Our context encoder is designed to be trained over all (or window) of past utterances



Our context encoder is designed to be trained over all (or window) of past utterances



Overview

- ✓ ❑ How to preserve and integrate "conversational context"?
- ✓ ❑ How to encode "conversational context"?
- ✓ ❑ **Experiments and Analysis**

Experiments

End-to-End ASR implemented in PyTorch based on ESPnet w/ word-level output

Acoustic Features

- 80d Filterbank + 3d Pitch (without deltas)

Encoder/ Decoder Characteristics

- Encoder: CNN (downsampled to 1/4) +BLSTM (6-layer 320 cells) – plus CTC
- Decoder: LSTM (2-layer 300 cells)

Decoding

- Beam search with width 10 (without external LM)

https://github.com/espnet/espnet

300 hours of Switchboard task

Switchboard task: two speaker conversations over telephone

	training	validation	evaluation	
	SWBD	SWBD	SWBD	CallHm
Conv.	$\boxed{2,402}$	34	20	20
Utters./Conv.	80	118	92	131

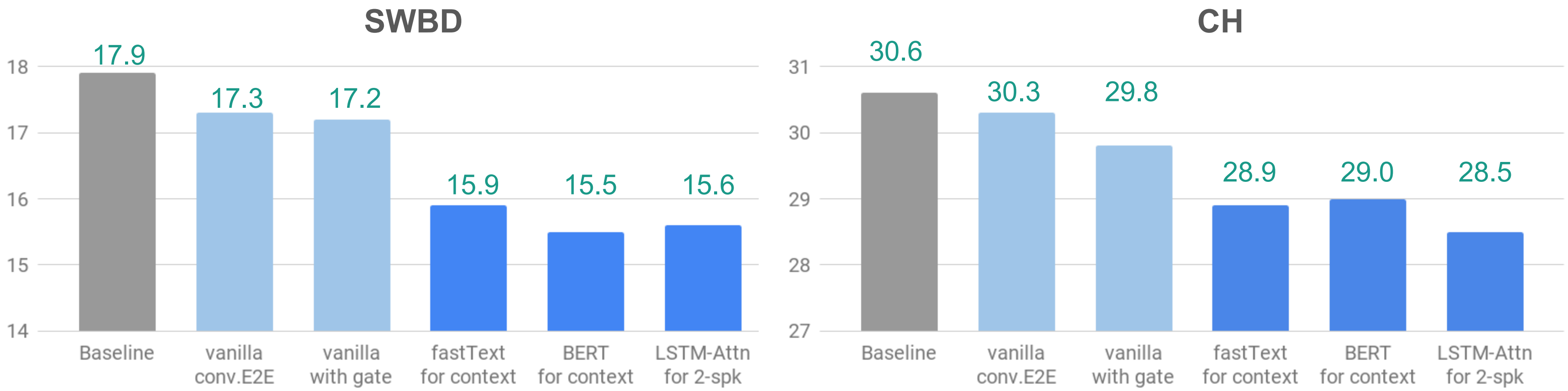
Related work and our baseline results

Our systems (no data augment.) are relatively small to train faster & fit GPU better

Model		#params.	LM	SWB	CH
Other E2E systems					
CTC (Zweig et al., 2017)	Char output	53M	✓	19.8	32.1
CTC (Audhkhasi et al., 2017)	Word output, phone pretrain	n/a	✗	14.6	23.6
Seq2Seq (Zeyer et al., 2018)	BPE-1k, layer-wise pretrain	*150M	✗	13.1	26.1
LF-MMI (Hadian et al., 2018)	Char output, data augment.	26M	✓	13.0	23.6
Seq2Seq (Park et al., 2019)	BPE-1k, data augment.	360M	✗	7.2	14.6
Our baseline					
our baseline	Char output	23M	✗	19.0	34.4
our baseline	Word-10k output	32M	✗	17.9	30.6

Our conv. E2E model outperforms over baseline

WER over different proposed context encoder methods



Kim et al, in ACL 2019
Kim et al, in Interspeech 2019

Updates on SWBD 300 hours task

Using BPE-1k output unit shows better performance than Word-10k

Model		#params.	LM	SWB	CH
Our baseline					
our baseline	Word-10k output	32M	✗	17.9	30.6
our baseline	BPE-1k output	24M	✗	15.0	28.1
Our conv. E2E					
our conv. E2E	BPE-1k output	25M	✗	14.4	27.5

Our conv. E2E model is also effective on other large datasets - including 2,000 hours of Fisher

Fisher has 11.7 k conversations

Model		#params.	LM	SWB	CH
Other E2E systems					
CTC (Zweig et al., 2017)	Char output	n/a	✓	10.2	17.7
CTC (Audhkhasi et al., 2018)	Word output, phone pretrain	n/a	✗	8.8	13.9
LF-MMI (Hadian et al., 2018)	Char output, data augment.	26M	✓	12.0	21.9
Seq2Seq (Battenberg et al., 2017)	Char output	120M	✗	8.6	17.8
Seq2Seq (Weng et al., 2018)	Char output, MBR	n/a	✗	8.3	15.5
Our systems					
our baseline	BPE-1k output	24M	✗	9.5	17.3
our conv. E2E	BPE-1k output	25M	✗	9.3	16.7

3,700 hours of medical conversations between doctor and patient

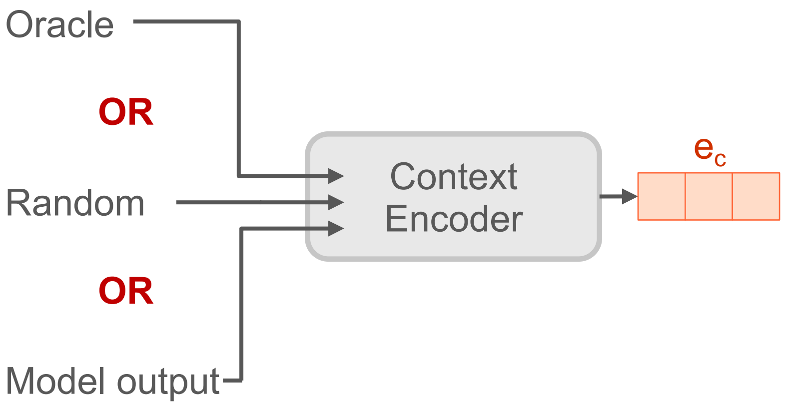
	training	validation	evaluation
	Medical	**Medical**	**Medical**
Conv.	25,500	45	100
Utters./Conv.	155	149	151

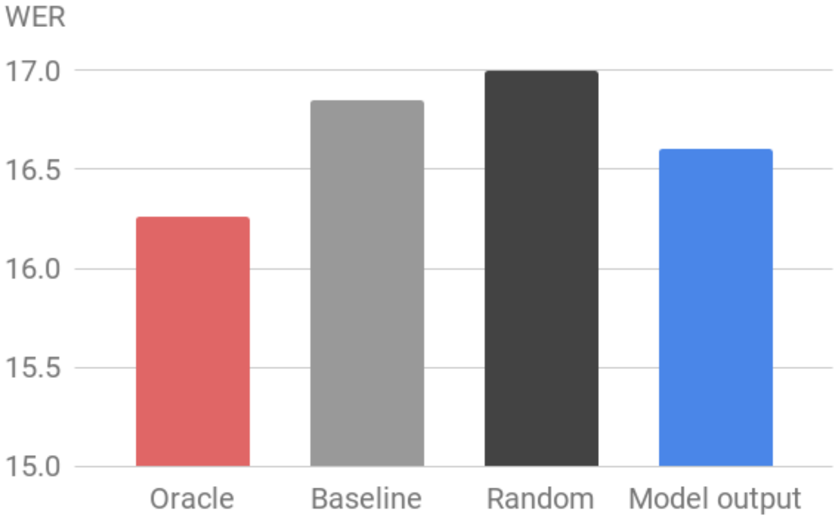
Model		#params.	LM	Medical
Our systems				
our baseline	BPE-1k output	24M	✗	22.1
our conv. E2E	BPE-1k output	25M	✗	21.6

This dataset is from UPMC from Pittsburgh hospital, which is unique, not publicly available, so there is no other benchmark results.

We validate the effect of context by comparing Oracle / Random Performance

Use oracle, random, model outputs, during decoding to study influence.

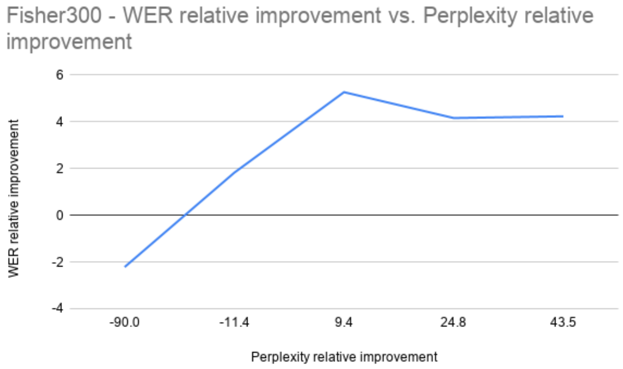


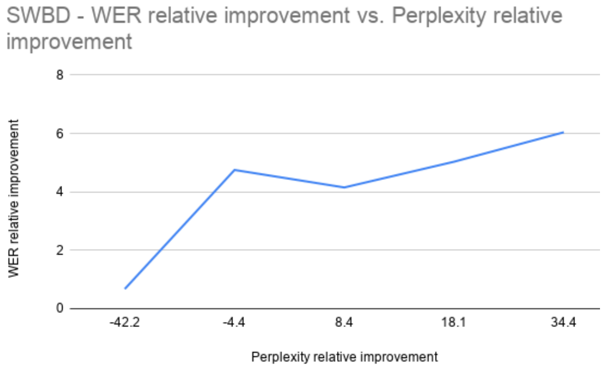


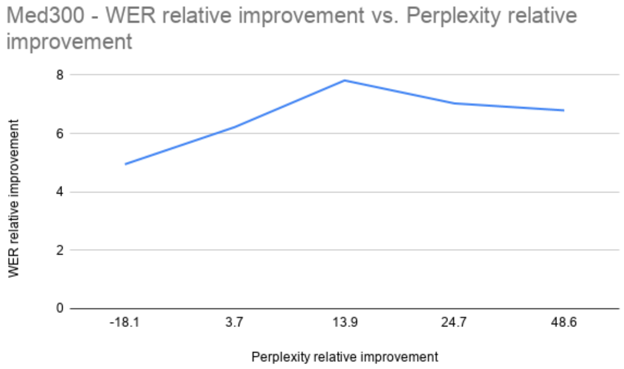
Kim et al, NAACL 2019

We validate the effect of context by comparing Perplexity improvement vs. WER improvement

- I used the result of LM and split utterances into 5 chunks in its improvement of context LM (ruled out AM). Then, I checked WER improvements of context ASR for each chunk of utterances



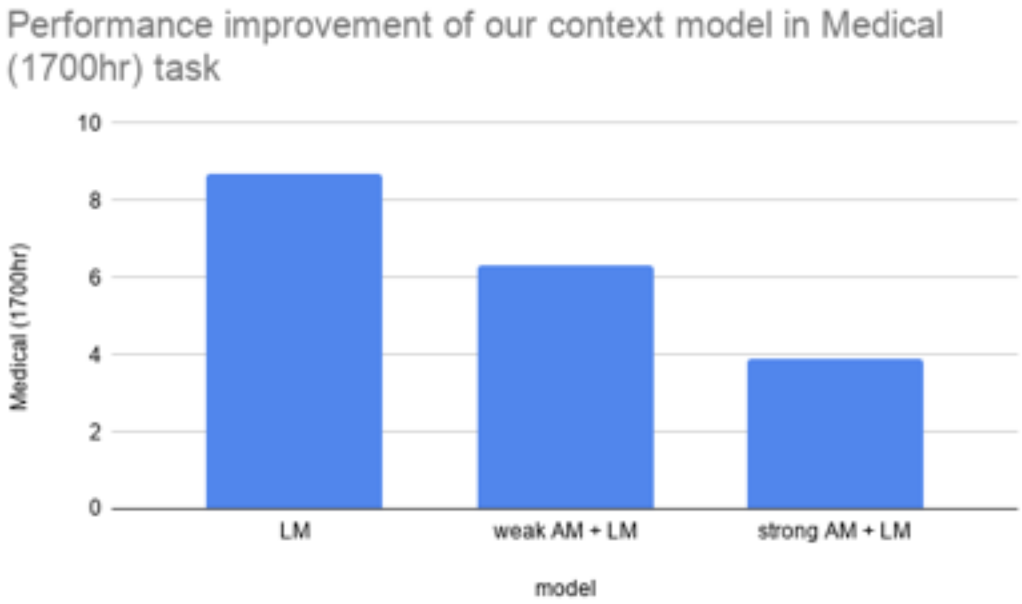




➔ Perplexity improvement of context LM translates to WER improvement of context ASR

How does the strength of AM affect the effectiveness of our linguistic context model?

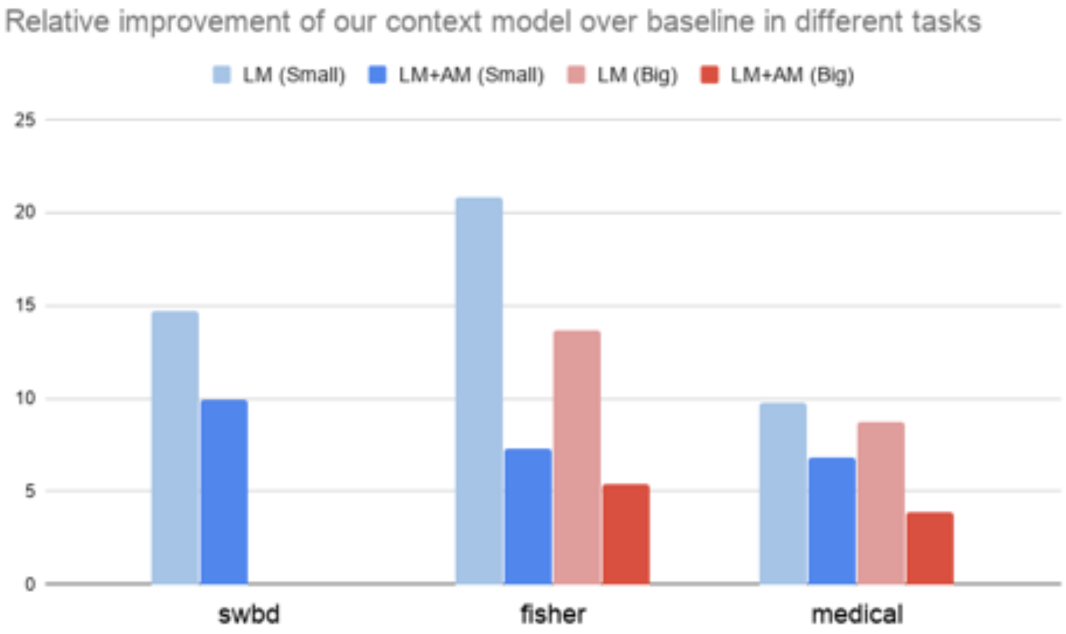
- I built LM and context LM and checked the relative improvement of perplexity (completely ruled out the AM)
- I built ASR with weak AM by reducing the encoder layer from 6 to 1



➔ Our linguistic context model performs less effectively as the AM gets stronger

How do large & small training datasets affect our context model?

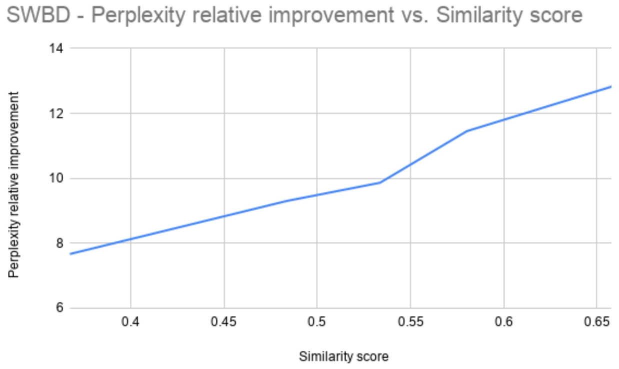
- ➢ I made Fisher/Medical training datasets in a size similar to SWBD
- ➢ I observed improved benefit of our context models in small datasets (blue bars)

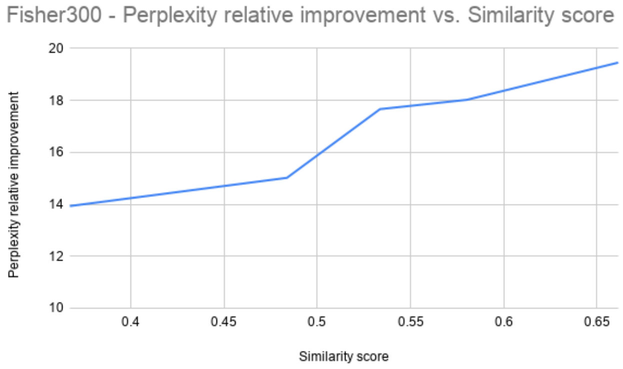


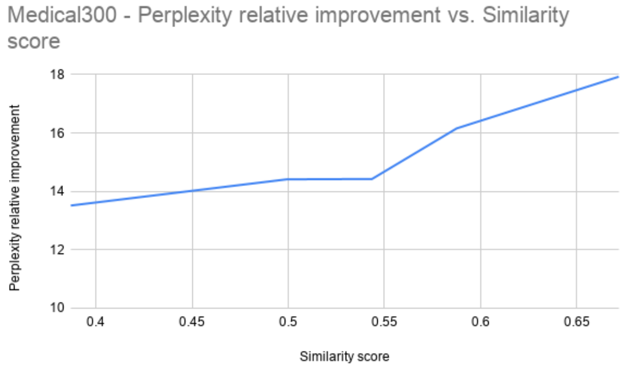
➔ Our linguistic context model perform less effectively with a large training dataset due to the strength of the AM from the large training dataset

How does our context model work? (1/3)

- The similarity score of an utterance (X-axis) is mean of cosine similarity of current utterance and the [1-10] historical utterances
- To get a single vector for each utterance, I use average of each output token from external pretrained LM (BERT)



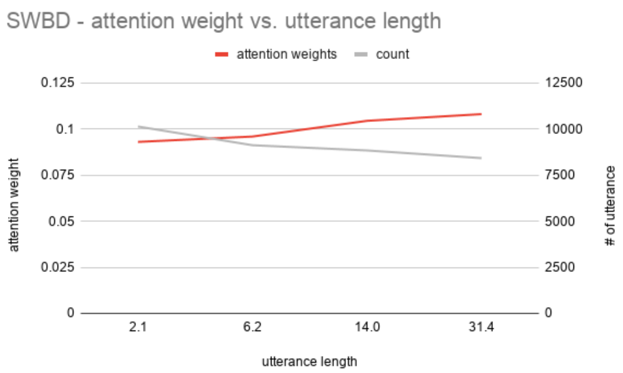


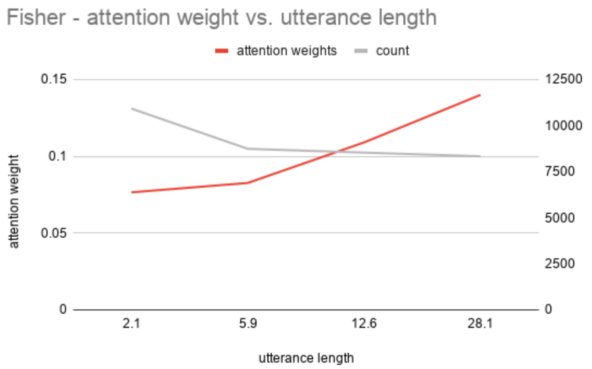


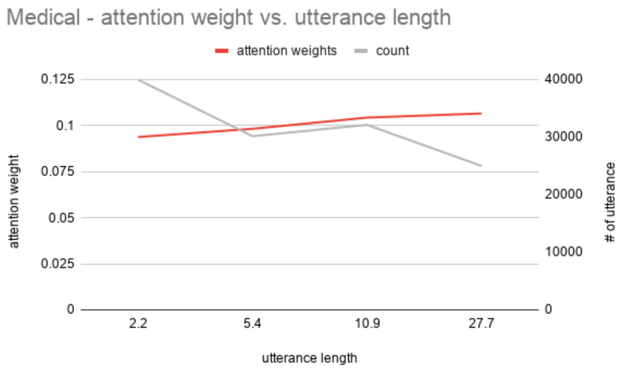
➔ Our context model performs better when historical utterances and current, predicted utterance are similar

How does our context model work? (2/3)

- I grouped historical utterances based on length of them, and checked the each group’s mean of length, mean of attention weight, and # of utterances
- The attention weight over the [10] historical utterances



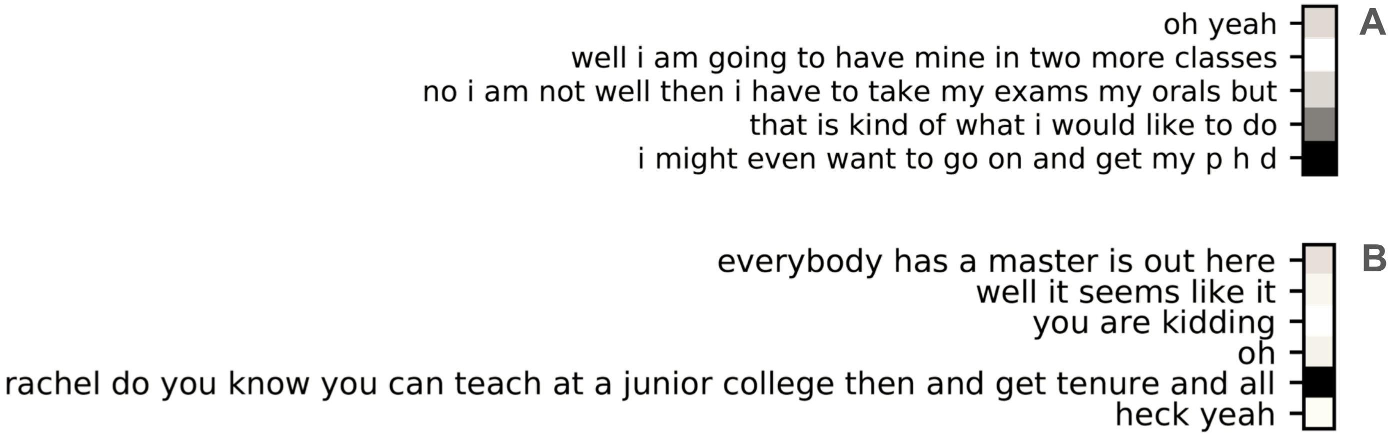




➔ Our context model tends to attend a long, more informative utterances

How does our context model work? (3/3)

Attention over utterance history of speaker A (top, “other”) and speaker B (bottom, “self”). Dark color represents higher weights.



Prediction of utterance for B: “come out here to California”

➔ **Our context model tend to attend a long, more informative utterances**

Kim et al, Interspeech 2019

Why reduced performance in medical task?

- Our context model is currently using two external pretrained LM: 1) BERT, and 2) fastText
- In case of medical task, out-of-vocabulary (OOV) rate of fastText is much higher than eval2000 (SWBD/Fisher) task

Test set	Eval2000	Medical
Total # of words	39,265	127,673
# of oov	1,361	8,316
oov rate	3.4%	6.5%

➔ Our context model doesn't fully take advantage of one of external pretrained LM (fastText) due to the high OOV rate in medical task

Cherry-picked examples from SWBD task (1/2)

Repeated word in current prediction may benefit from our model (e.g., utterance history)

Reference	*Baseline*	*Conv.E2E*
oh boy you guys been all over ***you guys** been all over*	*oh boy you guys been all* ***he goes** been all over*	*oh boy you guys been all* ***you guys** have been all over*
0.118	*0.172*	*0.169*
the the name does not really matter *okay we- my point is the experience* *uh-huh* *you know and specialization* *if if i have **specialization** in three areas*	*than the name was really not can't* *well my point is the uh experience* *uh-huh* *and special* *if i have **fushaliziation** in three air yes*	*then the name was really natural okay* *well my point is is the uh experience* *uh* *and specializat* *if if i have **specialization** in three areas*
0.106	*0.111*	*0.107*

Cherry-picked examples from SWBD task (2/2)

Semantically related word in current prediction may benefit from our model (e.g., utterance history)

Reference	*Baseline*	*Conv.E2E*
*i mean he was not proficient at it like doctor clausen is so he just put it in the muscle and figured it will i- it will get somewhere near the joints but it is not the same as when you put it in the **joints***	*i mean he wasn't professional at it like doctor classines so he just put it in the model and figured it will it it'll get somewhere near the georgia but it's not the same as when you put it in the **john***	*i mean he wasn't proficient at it like doctor closson is so he just put it in the muscle and figured it'll it it'll get somewhere near the joints but it is not the same as than when you put it in the **joint***
0.130	*0.142*	*0.138*

Cherry-picked examples from medical task (1/3)

Medical word in current prediction may benefit from our model (e.g., utterance history, “world knowledge”)

Reference	*Baseline*	*Conv.E2E*
all day had a fever then i don’t think i had a fever by the time i came in on thursday but motrin sometimes ***uh naprosyn*** *can mask*	*all the had a fever but i don’t think i had a fever by the time i it came and on thursday but motrin sometimes* ***an approsyn*** *can mass*	*all the had a fever then i don’t think i had a fever by the time i it came in on thursday but motrin sometimes* ***a naprosyn*** *can mass*
0.110	*0.161*	*0.141*

Cherry-picked examples from medical task (2/3)

Medical word in current prediction may benefit from our model (e.g., utterance history, “world knowledge”)

Reference	*Baseline*	*Conv.E2E*
it’s the uh it comes in a little tube it’s like a cream *like a white uh white cream.* *uh do you ever get improvement with the ultraviolet light at all* *uh not really* *like summer time the* ***psoriasis*** *doesn’t do great if you’re outside*	*it’s the it comes in a little tube it’s like a cream* *like a white uh.* *um do you ever get improvement with ultraviolet lights at all* *uh not really* *like summertime* ***with the rise it*** *doesn’t do grade if you’re outside*	*it’s the it comes in a little tube it’s like a cream* *like a white uh* *um do you ever get improvement with ultraviolet lights at all* *uh not really* *like summertime the* ***psoriasis*** *doesn’t do great if you’re outside*
0.098	*0.128*	*0.115*

Cherry-picked examples from medical task (3/3)

Medical word in current prediction may benefit from our model (e.g., utterance history, “world knowledge”)

Reference	*Baseline*	*Conv.E2E*
better to help more more control okay because again i think your brain looks pretty .. they’re more the same than they are different but i kind have the idea that more severe disease we should use one more ..once a week avonex now ***plegridy*** *is what is kind of preferred*	*better to help them more control okay because again i think your your brain looks pretty .. they’re more the same than there are different but i’m kind of out the idea that more severe disease you should use one more ..once a week avonex now* ***plaguery*** *is what is kind of preferred*	*better to help them more control okay because again i think your your brain looks pretty .. they’re more the same than there are different but uh kind of at the idea that more severe disease you use one more ..once a week avonex now* ***plegridy*** *is what is kind of preferred*
0.120	*0.163*	*0.157*

Conclusions

- We present an effective way to process conversations in end-to-end manner, rather than isolated utterances
- How to preserve and integrate “Context”?
  - Data serialization, Gated contextual decoder
- How to encode “Context”?
  - Context encoder with ”world knowledge”: BERT, speaker turn info
- Experiments and Analysis
  - Improved WER as well as conversational similarity

Future Work

- Improving baseline performance through tuning & bigger models
- Improving context representation
- Including “**acoustic**” conversational context in addition to “linguistic” conversational context
    - Emotions, speaking styles, background noise, non-verbal cues …
- Our approach can be potentially applied to,

  tasks from **long audio** to NLU, slot-filling, actions, summarization, IR, QA, …

Acknowledgement

Thesis Committee

- Florian Metze (Advisor) - CMU
- Richard M. Stern (Co-advisor) - CMU
- Bhiksha Raj - CMU
- Mike Seltzer - Facebook
- Shinji Watanabe - JHU

Collaborators

- Takaaki Hori - MERL
- Siddharth Dalmia – CMU PhD student

Fellowships & Funding

Center for Machine Learning and Health
Carnegie Mellon University

UPMC
LIFECHANGINGMEDICINE

SONY

References

Xiong et al, “The Microsoft 2017 Conversational Speech Recognition System”, in ICASSP 2018

Liu et al, “Dialog context language modeling with recurrent neural networks”, in ICASSP 2017

Pundak et al, “Deep context: end-to-end contextual speech recognition”, in SLT 2018

Chan et al, “Listen, attend, and spell: A neural network for large vocabulary conversational speech recognition”, in ICASSP 2016

Bahdanau et al, “End-to-end attention-based large vocabulary speech recognition”, in ICASSP 2016

Bojanowski et al, “Enriching word vectors with subword information”, in TACL 2017

Joulin et al, “Fasttext.zip: Compressing text classification models”, in arXiv 2016

Devlin et al, “BERT: Pre-training of deep bidirectional transformers for language understanding”, in NAACL 2019

Paszke et al, “Automatic differentiation in pytorch”, in NIPS workshop 2017

Watanabe et al, “ESPnet: End-to-End Speech Processing Toolkit”, in Interspeech 2018

References (2)

Zweig et al, “Advances in all-neural speech recognition”, in ICASSP 2017

Audhkhasi et al, “Direct Acoustics-to-Word models for English conversational speech recognition”, in Interspeech 2017

Zeyer et al, “Improved training of end-to-end attention models for speech recognition”, in Interspeech 2018

Hadian et al, “End-to-end speech recognition using lattice-free MMI”, in Interspeech 2018

Park et al, “SpecAugment: A simple data augmentation method for automatic speech recognition”, in Interspeech 2019

Sennrich et al, “Neural machine translation of rare words with subword units”, in ACL 2016

Battenberg et al, “Exploring neural transducers for end-to-end speech recognition”, in ASRU 2017

Weng et al, “Improving attention based sequence-to-sequence models for end-to-end English conversational speech recognition” in Interspeech 2018

Chiu et al, “speech recognition for medical conversations”, in Interspeech 2018

Publications

Suyoun Kim, Siddharth Dalmia, & Florian Metze, “Cross-Attention End-to-End ASR for Two-Party Conversations”, in INTERSPEECH 2019

Suyoun Kim, Siddharth Dalmia, & Florian Metze, “Gated Embeddings in End-to-End Speech Recognition for Conversational-Context Fusion”, in ACL 2019

Suyoun Kim, & Florian Metze, “Acoustic-to-Word Models with Conversational Context Information”, in NAACL 2019

Suyoun Kim*, Siddharth Dalmia*, & Florian Metze, “Situation Informed End-to-End ASR for CHiME-5 Challenge”, in SLT 2018

Suyoun Kim, & Florian Metze, “Dialog-context Aware End-to-End Speech Recognition”, in SLT 2018

Suyoun Kim, et. al, “Improved Training for Online End-to-End Speech Recognition”, in INTERSPEECH 2018

Suyoun Kim, & Michael L. Seltzer, “Towards Language-universal End-to-End Speech Recognition”, in ICASSP 2018

Publications (2)

<u>Suyoun Kim</u>, Takaaki Hori, & Shinji Watanabe, "Joint CTC-Attention based End-to-End Speech Recognition using Multi-task Learning", in ICASSP 2017

<u>Suyoun Kim</u>, & Ian Lane, "Recurrent Models for Auditory Attention in Multi-Microphone Distant Speech Recognition", in INTERSPEECH, 2016

<u>Suyoun Kim</u>, & Ian Lane, "Recurrent Models for Auditory Attention in Multi-Microphone Distant Speech Recognition" (earlier version), in ICLR workshop, 2016

<u>Suyoun Kim</u>, Bhiksha Raj, & Ian Lane, "Environmental Noise Embeddings for Robust Speech Recognition", in arXiv, 2016

Thank you!
Any Questions?

Suyoun Kim

suyoun@cmu.edu

http://suyoun.kim

Appendix