Deliberation Model Based Two-Pass End-to-End Speech Recognition

Ke Hu, Tara N. Sainath, Ruoming Pang, Rohit Prabhavalkar

ICASSP 2020

FR3.PB: Large Vocabulary Continuous Speech Recognition and Search
Session Type: Poster
Time: Friday, 8 May, 15:15 - 17:15
Location: Poster Area B

Outline

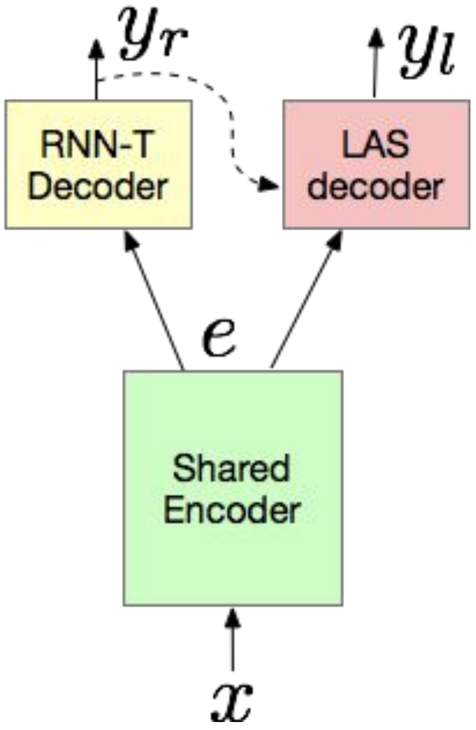
- Background
- Model Architecture
- Training and Decoding
- Experimental Analysis
- Comparison
- Conclusion

Outline

- Background
- Model Architecture
- Training and Decoding
- Experimental Analysis
- Comparison
- Conclusion

Background

- LAS based 2-pass model attends to acoustics and shows the state-of-the-art results [1]
- Neural denorm shows positive results by attending to text alone [2]
- Can we combine the two?



LAS Rescroing Model

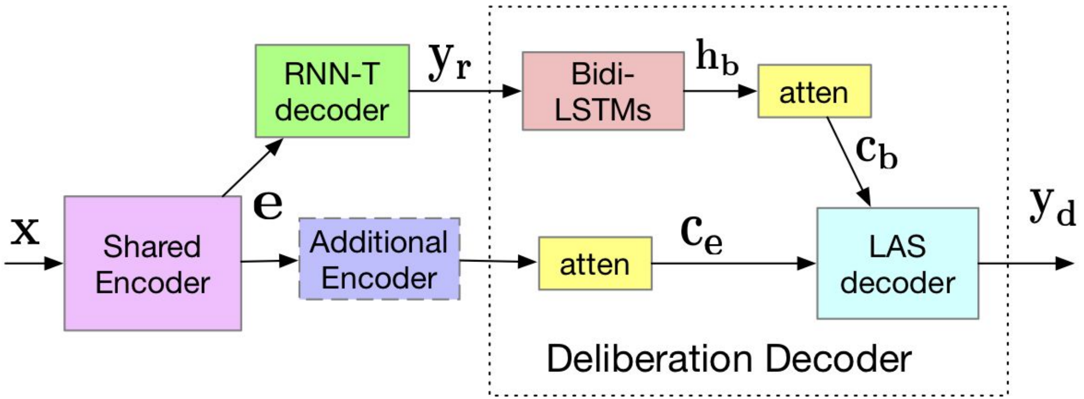
[1] Sainath et. al., Two-pass end-to-end speech recognition. Proc. Interspeech'19
[2] Peyser et al., Improving performance of end-to-end ASR on numeric sequences. Proc. Interspeech'19

Outline

- Background
- Model Architecture
- Training and Decoding
- Experimental Analysis
- Comparison
- Conclusion

Deliberation Model* (Xia et al.’17)

- Attention on both acoustic embeddings and RNN-T hypotheses
- Training: Init enc/dec from 1-pass RNNT
- Typically beam search decoding, but we also explore rescoring



* Y. Xia, F. Tian, L. Wu, J. Lin, T. Qin, N. Yu, T. Y. Liu, “Deliberation networks: Sequence generation beyond one-pass decoding”, In Advances in Neural Information Processing Systems, pp. 1784-1794, 2017.

Google

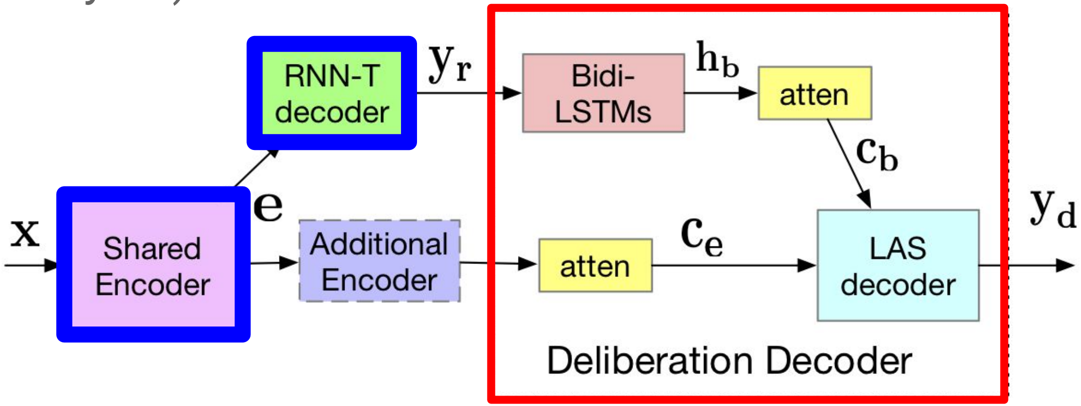
Outline

- Background
- Model Architecture
- Training and Decoding
- Experimental Analysis
- Comparison
- Conclusion

Training

Two step training:

1. Train the first-pass RNN-T model
2. Fix the RNN-T model and train the deliberation decoder (and possibly additional encoder layers)

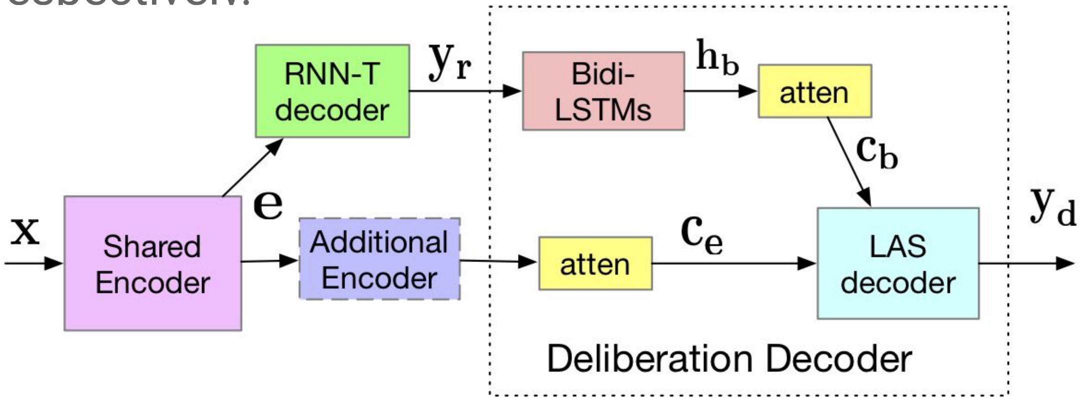


Joint Training

- An alternative to the second step in training is to train both RNN-T and deliberation decoder using a jointly loss:

$$L_{\text{joint}}(\theta_e, \theta_1, \theta_2) = L_{\text{RNNT}}(\theta_e, \theta_1) + \lambda L_{\text{CE}}(\theta_e, \theta_2)$$

- $\theta_e, \theta_1, \theta_2$: Parameters of shared encoder, RNN-T decoder, and deliberation decoder, respectively.



MWER-based Fine Tuning

- MWER training follows the previous two-step training to further reduce WER
  - Only update the deliberation decoder
- MWER Loss [Prabhavalkar et al.'18]:

$$L_{\mathrm{MWER}}(\mathbf{x}, \mathbf{y}^*) = \sum_{i=1}^{B} \hat{P}(\mathbf{y}_d^i | \mathbf{x})[W(\mathbf{y}_d^i, \mathbf{y}^*) - \hat{W}]$$

- In practice, we combine MWER loss with CE loss to stabilize training:

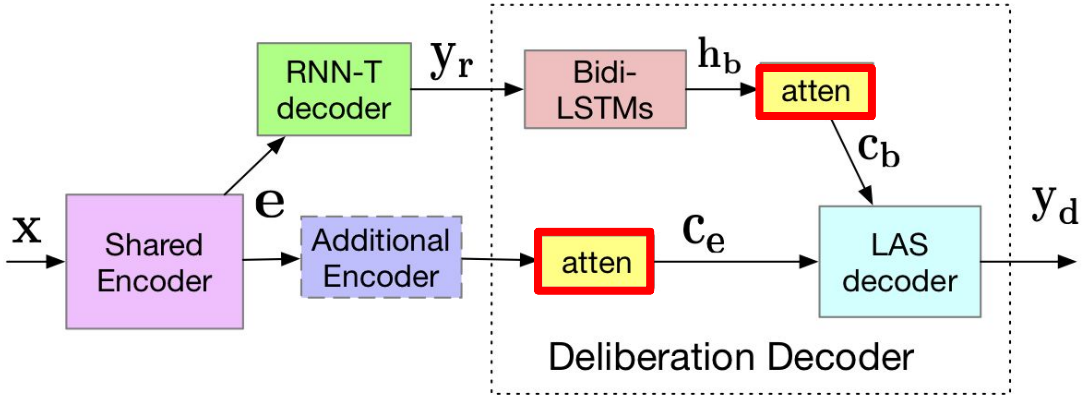
$$L'_{\mathrm{MWER}}(\mathbf{x}, \mathbf{y}^*) = L_{\mathrm{MWER}}(\mathbf{x}, \mathbf{y}^*) + \alpha L_{\mathrm{CE}}(\mathbf{x}, \mathbf{y}^*)$$

where $\alpha = 0.01$

Decoding

Two-step decoding:

1. Decode RNN-T to obtain the first-pass hypotheses $\mathbf{y}_r$
2. Attend to both encoded first-pass hypothese and encoder outputs
    a. Beam search decode
    b. Rescoring



Outline

- Background
- Model Architecture
- Training and Decoding
- Experimental Analysis
- Comparison
- Conclusion

Ablation Study: # RNN-T hypotheses

- Attend to different number of RNN-T hypotheses (pre-MWER)
- Did not learn the order of hypotheses → MWER training

ID	# RNN-T hyps	VS WER
E1	1-hyp	5.5
E2	2-hyp	5.4
E3	4-hyp	5.4
E4	8-hyp	5.4

pre-MWER does not improve much for VS

Improvement #1: MWER training

- Pre-MWER results in parentheses

ID	Models	VS WER
E1	1-hyp	5.4 (5.5)
E2	2-hyp	5.3 (5.4)
E3	4-hyp	5.2 (5.4)
E4	8-hyp	**5.1** (5.4)

MWER helps more when there are multiple hypotheses

Ablation Study: Acoustics Only or Text Correction

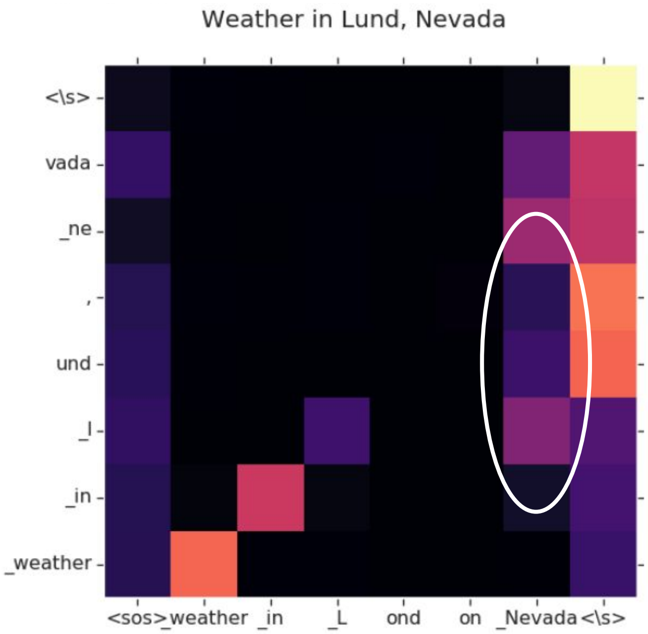
- Which attention is more useful (pre-MWER)

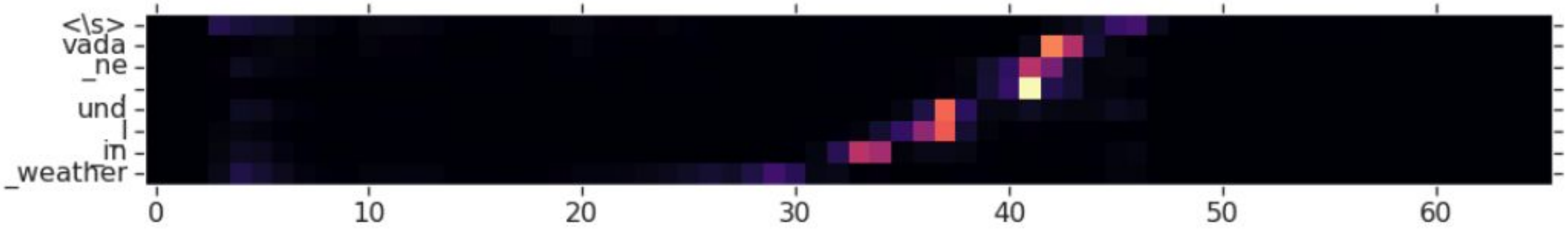
ID	Models	VS WER
B0	RNN-T Baseline	6.7
E5	Attend to acoustics alone	6.1
E6	Attend to 8 hyps alone	6.1
E4	Attend to both	5.4

Acoustic and text information is complementary

Attention Plots

- Target: Weather in Lund, Nevada
  - Top RNN-T hyp: Weather in London Nevada
- Attention on RNN-T hypotheses look ahead for context
- Simultaneously focus on relevant acoustic frames





Improvement #2: Additional Encoder (pre-MWER)

- Additional encoder (AE) helps both LAS and deliberation models

ID	Model	VS WER (%)
E4	8-hyp Delib	5.4
E7	E4 + AE	5.2
B1	LAS	6.1
B2	LAS + AE	5.8

Deliberation Decoder as a Rescorer

- Rescoring using bidirectional encoding should help compared to LAS decoder
- Promising results since deliberation does not have AE

ID	Model (pre-MWER)	VS WER (%)
E8	RNN-T + LAS Rescoring (w/ AE)	6.0
B3	8-hyp Deliberation Rescoring	5.7

Deliberation model can also be used as a rescorer

Google

Improvement #3: Joint Training

- Jointly train RNN-T encoder & decoder, and deliberation decoder

Deliberation Model	VS WER (%)
8-hyp, post-EMBR	5.1
+ Joint training	5.0

- Improved RNN-T
  - VS WER: 6.7% (baseline RNN-T) → 6.4%

Outline

- Background
- Model Architecture
- Training and Decoding
- Experimental Analysis
- Comparison
- Conclusion

Model Comparison for VS

ID	Model	Decoding	VS WER (%)
B0	RNN-T	Beam Search	6.7
B4	LAS [10]	Rescoring	5.7
B5	LAS [10]	Beam Search	5.5
B9	Deliberation	Beam Search	5.1
E10	+ Joint training	Beam Search	**5.0**

Deliberation model improves in general by attending to RNN-T hypotheses

Proper Noun Test sets

- Deliberation model performs better on proper noun test sets
  - 16% better than LAS beam search on SxS set
  - 23% better than LAS rescoring on SxS set

Model	Decoding	WER (%)					Estimated GFLOPS
		SxS	Songs	Contacts-Real	Contacts-TTS	Apps	
RNN-T	Beam search	35.2	11.9	15.9	24.3	7.8	3.5
LAS [10]	Rescoring	31.4	10.9	14.7	22.6	7.5	4.8
LAS [10]	Beam search	**29.0**	**11.7**	**14.7**	**22.9**	**8.3**	**4.8**
Deliberation	Beam search	26.6	9.9	13.7	22.3	7.1	8.8
+ Joint training	Beam search	**24.3**	**9.6**	**13.4**	**22.0**	**6.4**	**8.8**

Computation Cost Comparison

- Deliberation model performs better on proper noun test sets
  - Estimate decoder computation cost by gigaFLOPS (GFLOPS)

$$\text{FLOPS} = M_B \cdot N \cdot H + M_D \cdot N \cdot B + \text{FLOPS}_{atten}$$

Model	Decoding	WER (%)					Estimated GFLOPS
		SxS	Songs	Contacts-Real	Contacts-TTS	Apps	
RNN-T	Beam search	35.2	11.9	15.9	24.3	7.8	3.5
LAS [10]	Rescoring	31.4	10.9	14.7	22.6	7.5	4.8
LAS [10]	Beam search	**29.0**	**11.7**	**14.7**	**22.9**	**8.3**	**4.8**
Deliberation	Beam search	26.6	9.9	13.7	22.3	7.1	8.8
+ Joint training	Beam search	**24.3**	**9.6**	**13.4**	**22.0**	**6.4**	**8.8**

Example Wins and Losses

- Wins: URLs, proper nouns, LM
- Losses: Spelling errors, over-correction of proper nouns

Ref	Deliberation	LAS Rescoring
quadcitytimes.com	quadcitytimes.com	Quality times.com
Walmart job application	Walmart job application	Where my job application
train near me	train near me	china near me
bio of Chesty Puller	bio of Chester Fuller	bio of Chesty Fuller
2016 Kia Forte5	2016 Kia Forte 5	2016 Kia Forte5

Outline

- Background
- Model Architecture
- Training and Decoding
- Experimental Analysis
- Comparison
- Conclusion

Conclusion

- Deliberation-based two-pass E2E model outperforms LAS rescoring in Google VoiceSearch and proper noun recognition in WER, by 12% and 23%, respectively
- The model also performs 21% relatively better than a large-scale conventional model for VoiceSearch
- The model needs more computation than LAS rescoring, and batching can improve latency