Adaptive Discounting of Implicit Language Models in RNN-Transducers

Vinit Unni [1 †], Shreya Khare [2 †], Ashish Mittal[2], Preethi Jyothi [1], Sunita Sarawagi[1], Samarth Bharadwaj[2]
[1]Indian Institute of Technology Bombay(India), [2]IBM Research (India), † Equal contribution.

Paper 8878

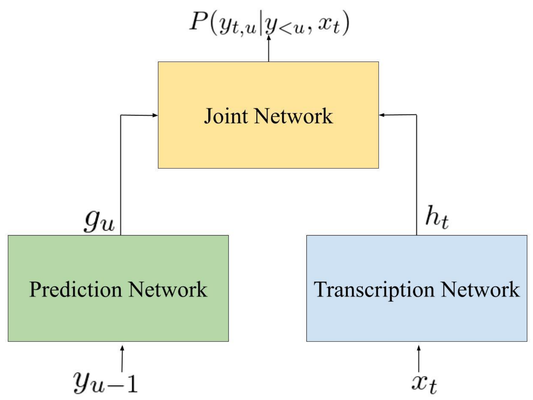
Motivation

- Fixing *hallucinated predictions* from RNN-Transducer models
- H*allucinated outputs* are acoustically inconsistent with the underlying speech
- Causes of errors:
  a. Domain mismatch between train and test
     Ex: REF: `Welcome to Canara Bank`
     HYP: `Welcome to Amazon`
  b. Memorisation of frequent training phrases
     Ex: REF: `Good morning ma'am`
     HYP: `Good evening ma'am`
- Indicative of LM bias

We propose AdaptLMD: A lightweight adaptive LM-discounting algorithm for RNN-T models

RNN-Transducers[1]

- Two encoder arms
  - *Transcription Network (*TN*)*: Generate acoustic representations
    $$h_t = TN(x, t)$$
  - *Prediction Network* (PN): Generate textual representations like an autoregressive LM
    $$g_u = PN(y_{<u})$$
- *Joint Network* (JN) combines both representations
  $$J(h_t \bigoplus g_u)$$
- Trained to maximize likelihood of the output text sequences by marginalizing over all possible alignments



Implicit Language/Acoustic Models

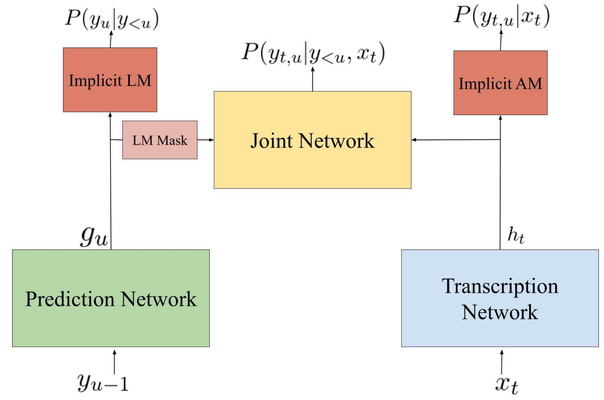
- Calculate implicit *AM/LM* [2] predictions
  $$P_{ILM}(y_u|y_{<u}) = softmax(J(0 \bigoplus g_u))$$
  $$P_{IAM}(y_t|x_t) = softmax(J(h_t \bigoplus 0))$$
- Training loss is RNN-T loss + *implicit LM/AM* loss

Adapt LMD

- Random LM Masking
  - Mask *PN* outputs during training to make the JN more robust to spurious implicit LM embeddings
- Token Rarity
  - Apply discounting on rare substrings
    $$P_{roll}(\mathbf{y} + y_u) = \rho P_{roll}(\mathbf{y}) + P_{ILM}(y_u)$$
- Discrepancy between AM and LM
  - Apply discounting where *AM* and *LM* disagree
    $$D(IAM||ILM) = KLD(P_{ILM}||P_{IAM})$$
- Final Scoring
  - RNNT score is discounted by implicit LM score

$$\tilde{S}_{\text{disc}}(y_{t,u}|\mathbf{y}, x_t) = \log P_{\text{rnnt}}(y_{t,u}|\mathbf{y}, x_t) - \lambda \max(0, D_{\text{adapt}}(y_{t,u}, \mathbf{y})) \log P_{\text{ILM}}(y_u|\mathbf{y})$$

$$D_{\text{adapt}}(y, \mathbf{y}) = \begin{cases} (1 - P_{\text{roll}}(\mathbf{y}))\mathrm{D}(P_{\text{ILM}}(y)||P_{\text{IAM}}(y)) & \text{if } y \neq \epsilon \\ 0 & \text{else.} \end{cases}$$



Examples

Ground truth	Baseline	AdaptLMD
Ki aap apanaa business account	*Ki aap apanaa* discount account	*Ki aap apanaa* business account
Ab aapki call transfer	*Ab aapki* block transfer	*Ab aapki* call transfer
Naam hai vicky rajak	Naam hai reeti raghav	Naam hai vikkee raaj
Ye online activate karavaa	mujhe online network karavaa	Ya online activate karavaa

Experiments and Results

- Numbers reported on a proprietary dataset
- Code-mixed speech (Hi-En) of 628 hours
- Banking (184.87 hrs), Insurance (165.08hrs), Retail (135.87 hrs) and Telco (142.33 hrs).

Test set	System	CER/WER	Rare CER/PER
Banking	Baseline	20.0/22.4	75.9/70.8
	AdaptLMD	**18.7/21.5**	**71.8/67.4**
Insurance	Baseline	17.7/18.4	76.7/70.2
	AdaptLMD	**16.4/17.7**	**67.2/60.7**
Retail	Baseline	22.4/24.7	81.3/76.7
	AdaptLMD	**21.2/23.7**	**78.3/72.2**
Telco	Baseline	18.6/19.4	72.6/68.3
	AdaptLMD	**17.5/18.6**	**68.6/64.3**

`Out-of-domain results`

Test Set	System	CER/WER	Rare CER/PER
All	Baseline	13.5/14.5	71.2/57.7
	AdaptLMD	**13.1/14.1**	**63.9/52.2**
Banking	Baseline	13.6/14.5	68.2/66.9
	AdaptLMD	**13.1/14.2**	**59.6/63.2**
Insurance	Baseline	14.4/15.0	68.2/66.9
	AdaptLMD	**14.0/15.1**	**59.6/63.5**
Retail	Baseline	13.2/14.2	72.7/65.3
	AdaptLMD	**12.9/13.8**	**63.8/63.1**
Telecom	Baseline	14.0/15.2	76.6/68.0
	AdaptLMD	**13.6/14.9**	**63.3/60.2**

`In-domain results`

Conclusions

- AdaptLMD can be used with any RNN-T and is dynamically invoked
- AdaptLMD leads to improvements on rare-word predictions and also consistently benefits overall WERs

References

[1] A. Graves: Sequence Transduction with Recurrent Neural Networks
[2] M. Zhong, K. Naoyuki, G. Yashesh, P. Sarangarajan, S. Eric, L. Liang, C. Xie, L. Jinyu, Gong, Yifan: Internal Language Model Training for Domain-Adaptive End-to-End Speech Recognition