Investigating Methods to Improve Language Model Integration for Attention-based Encoder-Decoder ASR Models

Mohammad Zeineldeen[1,2], Aleksandr Glushko[1], Wilfried Michel[1,2], Albert Zeyer[1,2], Ralf Schlüter[1,2], Hermann Ney[1,2]

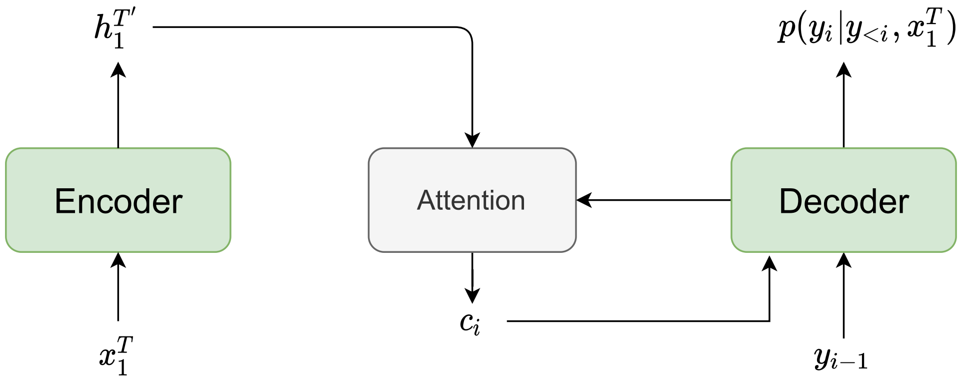
RWTH Aachen University[1], AppTek GmbH[2]
Interspeech 2021, Brno

Introduction

- Attention encoder-decoder (AED) models benefit from external language model integration

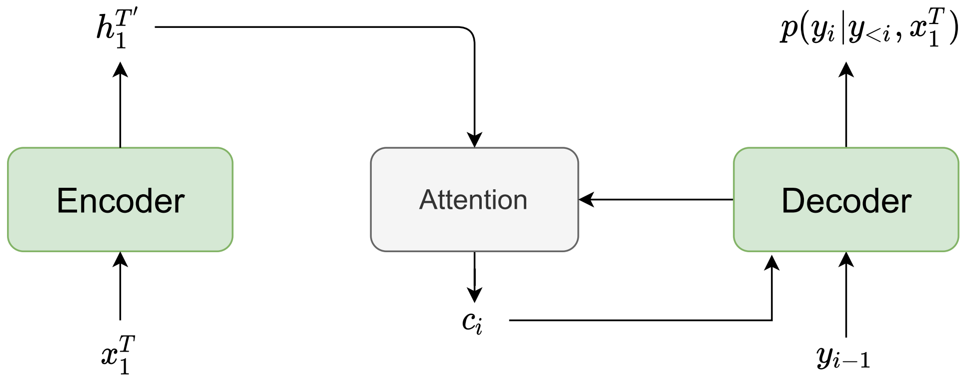
Introduction

- Attention encoder-decoder (AED) models benefit from external language model integration
- **Problem**: AED models learn an implicit **internal language model** (ILM) from the training data



Introduction

- Attention encoder-decoder (AED) models benefit from external language model integration
- **Problem**: AED models learn an implicit **internal language model** (ILM) from the training data



- How to compute the ILM probability for prior correction during recognition for better performance?

Internal Language Model Estimation

During recognition, the search algorithm searches for the best word sequence w_1^N that maximizes:

$$\hat{w}_1^{\hat{N}} = \underset{N, w_1^N}{\arg\max} \left\{ \log P(w_1^N | x_1^T) \right\}$$

Internal Language Model Estimation

During recognition, the search algorithm searches for the best word sequence w_1^N that maximizes:

$$\hat{w}_1^{\hat{N}} = \underset{N, w_1^N}{\arg\max} \left\{ \log P(w_1^N | x_1^T) \right\}$$

The posterior probability can be defined as:

$$P(w_1^N | x_1^T) \propto P_{\mathrm{AED}}(w_1^N | x_1^T) \cdot P_{\mathrm{LM}}^{\lambda_1}(w_1^N)$$

Internal Language Model Estimation

During recognition, the search algorithm searches for the best word sequence w_1^N that maximizes:

$$\hat{w}_1^{\hat{N}} = \underset{N, w_1^N}{\arg\max} \left\{ \log P(w_1^N | x_1^T) \right\}$$

The posterior probability can be defined as:

$$P(w_1^N | x_1^T) \propto P_{\mathrm{AED}}(w_1^N | x_1^T) \cdot P_{\mathrm{LM}}^{\lambda_1}(w_1^N) \cdot \boldsymbol{P_{\mathrm{ILM}}^{-\lambda_2}(w_1^N)}$$

Internal Language Model Estimation

During recognition, the search algorithm searches for the best word sequence w_1^N that maximizes:

$$\hat{w}_1^{\hat{N}} = \arg\max_{N, w_1^N} \left\{ \log P(w_1^N | x_1^T) \right\}$$

The posterior probability can be defined as:

$$P(w_1^N | x_1^T) \propto P_{\text{AED}}(w_1^N | x_1^T) \cdot P_{\text{LM}}^{\lambda_1}(w_1^N) \cdot \boldsymbol{P_{\text{ILM}}^{-\lambda_2}(w_1^N)}$$

The ILM is defined as:

$$P_{\text{ILM}}(w_1^N) = \sum_{T, x_1^T} P_{\text{AED}}(w_1^N | x_1^T) \cdot P(x_1^T)$$

Internal Language Model Estimation

During recognition, the search algorithm searches for the best word sequence w_1^N that maximizes:

$$\hat{w}_1^{\hat{N}} = \underset{N, w_1^N}{\arg\max} \left\{ \log P(w_1^N | x_1^T) \right\}$$

The posterior probability can be defined as:

$$P(w_1^N | x_1^T) \propto P_{\mathrm{AED}}(w_1^N | x_1^T) \cdot P_{\mathrm{LM}}^{\lambda_1}(w_1^N) \cdot \boldsymbol{P_{\mathrm{ILM}}^{-\lambda_2}(w_1^N)}$$

The ILM is defined as:

$$P_{\mathrm{ILM}}(w_1^N) = \sum_{T, x_1^T} P_{\mathrm{AED}}(w_1^N | x_1^T) \cdot P(x_1^T)$$

However, the summation is **intractable**.

Internal Language Model Estimation

During recognition, the search algorithm searches for the best word sequence w_1^N that maximizes:

$$\hat{w}_1^{\hat{N}} = \underset{N, w_1^N}{\arg\max} \left\{ \log P(w_1^N | x_1^T) \right\}$$

The posterior probability can be defined as:

$$P(w_1^N | x_1^T) \propto P_{\mathrm{AED}}(w_1^N | x_1^T) \cdot P_{\mathrm{LM}}^{\lambda_1}(w_1^N) \cdot \boldsymbol{P_{\mathrm{ILM}}^{-\lambda_2}(w_1^N)}$$

The ILM is defined as:

$$P_{\mathrm{ILM}}(w_1^N) = \sum_{T, x_1^T} P_{\mathrm{AED}}(w_1^N | x_1^T) \cdot P(x_1^T)$$

However, the summation is **intractable**.

$\rightarrow$ We propose different novel methods to estimate the ILM for AED models

Approches

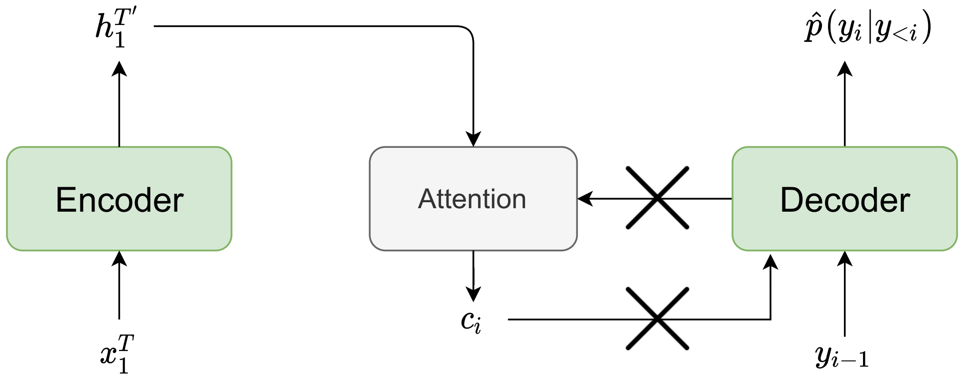
- ILM estimation methods can be classified as:

Approches

- ILM estimation methods can be classified as:
  1. Model-agnostic methods (e.g Density Ratio [McDermott & Sak+ 19])

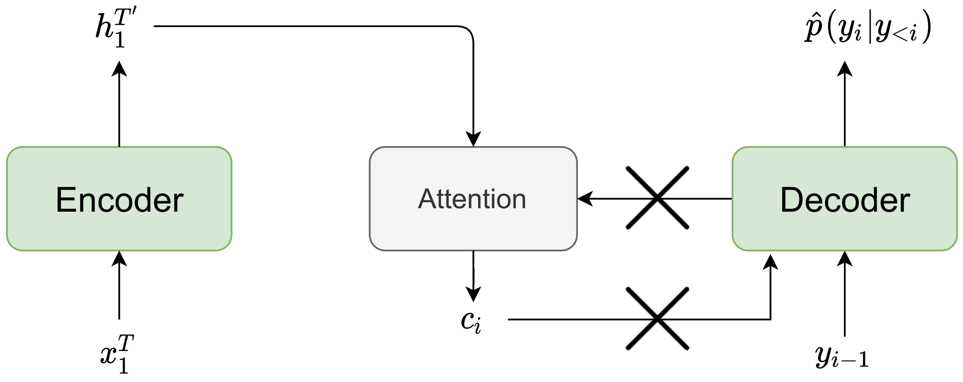
Approches

- ILM estimation methods can be classified as:
  1. Model-agnostic methods (e.g Density Ratio [McDermott & Sak+ 19])
  2. Model-specific methods [Variani & Rybach+ 20, Meng & Parthasarathy+ 20]



Approches

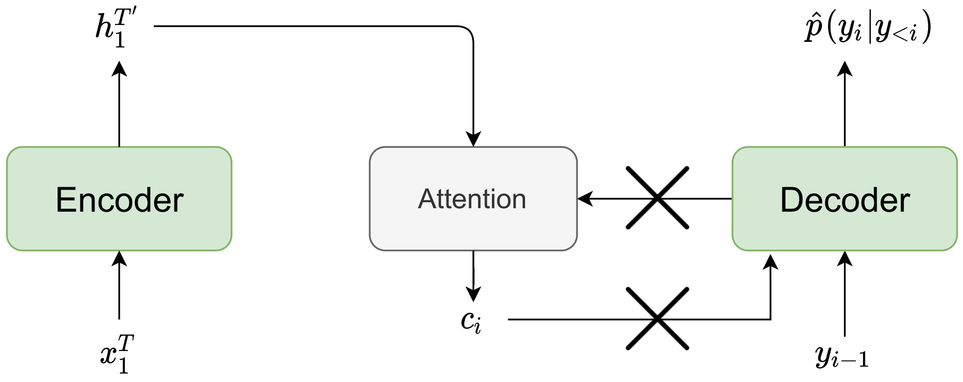
- ILM estimation methods can be classified as:
  1. Model-agnostic methods (e.g Density Ratio [McDermott & Sak+ 19])
  2. Model-specific methods [Variani & Rybach+ 20, Meng & Parthasarathy+ 20]



- We argue that using **encoder bias** can be helpful and this is more consistent with training

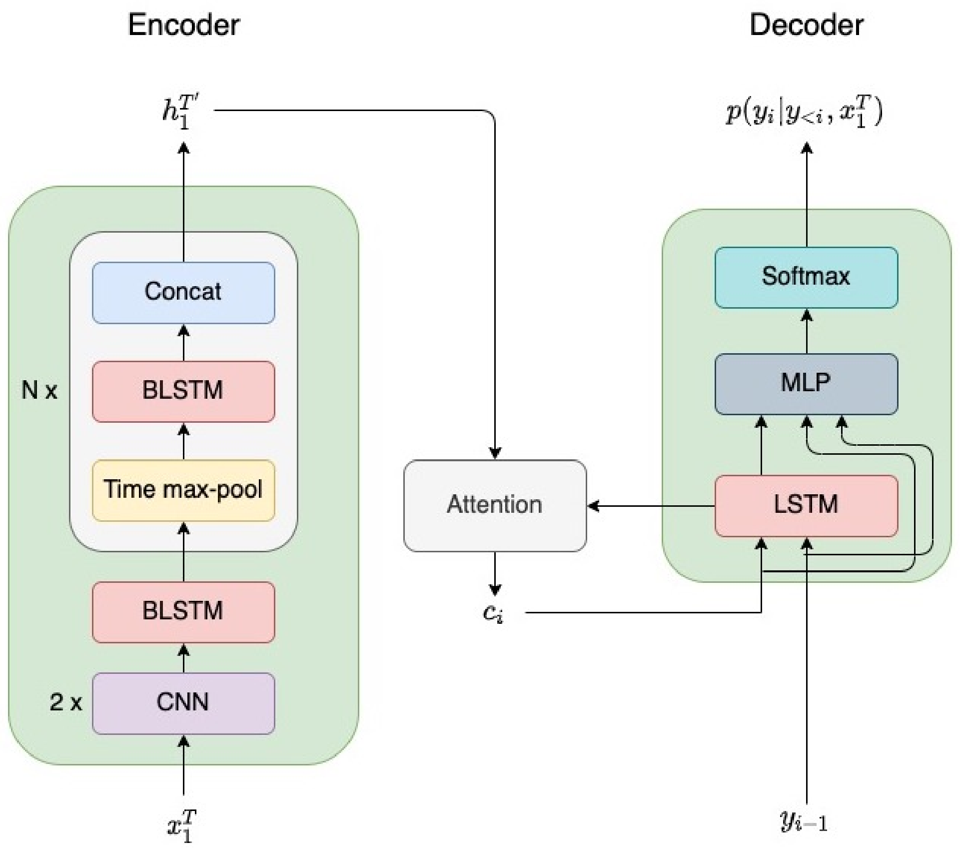
Approches

- ILM estimation methods can be classified as:
  1. Model-agnostic methods (e.g Density Ratio [McDermott & Sak$^+$ 19])
  2. Model-specific methods [Variani & Rybach$^+$ 20, Meng & Parthasarathy$^+$ 20]

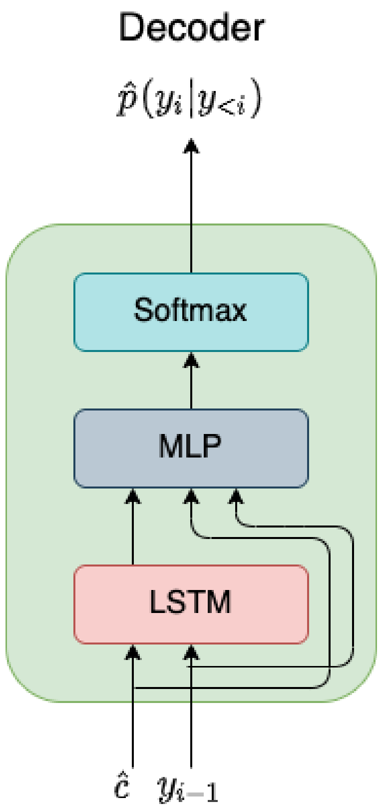


- We argue that using **encoder bias** can be helpful and this is more consistent with training
- This work focuses on **model-specific** estimation methods by replacing attention context vector with either static or trained context vectors

Attention Encoder-Decoder Model



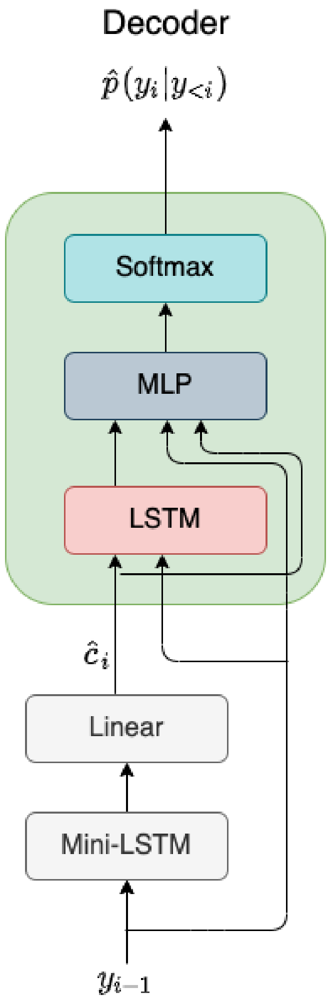
Static Context Vector Estimation



- Static vector $\rightarrow$ position independent
- Replace original context vector c_i by $\hat{c}$:
  - Zero vector (all elements are zero)
  - **Average** of all encoder states over train data
  - **Average** of all context vectors over train data

Trained Context Vector Estimation

- Training Steps
  1. **Freeze** all the parameters of AED model
  2. Add Linear and Mini-LSTM **trainable** layers
  3. Retrain the AED model for few epochs
- **Minimizes** directly the **perplexity**
- Trained only on transcription



ILM Suppression

- Limited Context Decoder
  - Replace the LSTM in the decoder with feed-forward layers
  - Less effective ILM
- Train AED together with LM via sequence training or local log-linear combination [Michel & Schlüter^{+} 20]
  - ASR model relies on the LM for language modeling and focuses on acoustic modeling

Results on Switchboard 300h

Method	WER [%]	
	Hub5'01	RT03
None	13.4	16.3
Shallow Fusion	13.0	15.7
Density Ratio	12.7	15.3
zero	12.9	15.6
$\mathbb{E}_{\mathcal{D}}[h]$	12.3	15.0
$\mathbb{E}_{\mathcal{D}}[c]$	12.4	14.9
$\mathbb{E}_{x}[h]$	12.6	15.2
Mini-LSTM	**12.2**	**14.8**

- ILM estimation by replacing attention context vector by:
  - zero: zero vector
  - $\mathbb{E}_{\mathcal{D}}[h]$: average of encoder states over train data
  - $\mathbb{E}_{\mathcal{D}}[c]$: average of context vectors over train data
  - $\mathbb{E}_{x}[h]$: average encoder states during recognition
  - Mini-LSTM: trained context vector
- Achieved **6% relative improvement** in terms of WER compared to Shallow Fusion

Results on LibriSpeech 960h

Method	WER [%]	
	dev-other	test-other
None	10.37	10.88
Shallow Fusion	6.80	7.59
Density Ratio	6.68	7.22
train w. LM	6.19	6.81
zero	6.43	6.96
$\mathbb{E}_{\mathcal{D}}[h]$	6.19	6.76
$\mathbb{E}_{\mathcal{D}}[c]$	6.19	6.74
$\mathbb{E}_{x}[h]$	6.34	7.01
Mini-LSTM	**5.76**	**6.53**

- train w. LM: train AED model with LM to suppress ILM
- ILM estimation by replacing attention context vector by:
  - zero: zero vector
  - $\mathbb{E}_{\mathcal{D}}[h]$: average of encoder states over train data
  - $\mathbb{E}_{\mathcal{D}}[c]$: average of context vectors over train data
  - $\mathbb{E}_{x}[h]$: average encoder states during recognition
  - Mini-LSTM: trained context vector
- Achieved **15% and 16% relative improvement** in terms of WER compared to Shallow Fusion

Cross-domain Evaluation

- ASR model trained on LibriSpeech 960h dataset
- Evaluated on TED-LIUM-V2 [Rousseau & Deléglise+ 14] dev and test datastes

Method	WER [%]	
	TLv2-dev	TLv2-test
None	22.0	22.9
Shallow Fusion	18.5	19.3
Density Ratio	16.6	17.8
zero	17.3	18.3
$\mathbb{E}_{\mathcal{D}}[h]$	16.7	17.5
$\mathbb{E}_{\mathcal{D}}[c]$	16.8	18.0
$\mathbb{E}_{x}[h]$	16.7	18.0
Mini-LSTM	**16.1**	**16.9**

- ILM estimation by replacing attention context vector by:
  - zero: zero vector
  - $\mathbb{E}_{\mathcal{D}}[h]$: average of encoder states over train data
  - $\mathbb{E}_{\mathcal{D}}[c]$: average of context vectors over train data
  - $\mathbb{E}_{x}[h]$: average encoder states during recognition
  - Mini-LSTM: trained context vector

Limited Context Decoder - Switchboard 300h

<table>
<tr><th rowspan="2">Method</th><th colspan="2">WER [%]</th></tr>
<tr><th>Hub5'01</th><th>RT03</th></tr>
<tr><td>None</td><td>14.0</td><td>16.8</td></tr>
<tr><td>SF</td><td>13.2</td><td>15.6</td></tr>
<tr><td>DR</td><td>13.2</td><td>15.6</td></tr>
<tr><td>zero</td><td>12.6</td><td>15.0</td></tr>
<tr><td>$\mathbb{E}_{\mathcal{D}}[h]$</td><td rowspan="2">12.4</td><td>14.8</td></tr>
<tr><td>$\mathbb{E}_{\mathcal{D}}[c]$</td><td rowspan="2">14.9</td></tr>
<tr><td>$\mathbb{E}_{x}[h]$</td><td>12.5</td></tr>
<tr><td>Mini-LSTM</td><td>12.6</td><td>14.9</td></tr>
</table>

- ILM estimation by replacing attention context vector by:
  - zero: zero vector
  - $\mathbb{E}_{\mathcal{D}}[h]$: average of encoder states over train data
  - $\mathbb{E}_{\mathcal{D}}[c]$: average of context vectors over train data
  - $\mathbb{E}_{x}[h]$: average encoder states during recognition
  - Mini-LSTM: trained context vector
- 1-layer FF decoder with context size 3
- **Average-based static** estimation methods perform better

Conclusions

- Subtracting the internal language model (ILM) during recognition gives significant improvements in terms of WER
- We proposed a novel method to train the attention context vector for ILM estimation which outperforms other methods
- We achieved 6% relative improvement in terms of WER on Switchboard 300h test sets as well as 15%-16% on LibriSpeech test sets
- Feed-forward or limited context decoder AED model can achieve comparable results to a recurrent decoder on Switchboard 300h task with ILM subtraction
- This work shows the importance of considering ILM subtraction in order to acheive better results

Thank you for your attention

Any questions?

References

[McDermott & Sak+ 19] E. McDermott, H. Sak, E. Variani.
A density ratio approach to language model fusion in end-to-end automatic speech recognition.
2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), Vol., pp. 434–441, 2019.

[Meng & Parthasarathy+ 20] Z. Meng, S. Parthasarathy, E. Sun, Y. Gaur, N. Kanda, L. Lu, X. Chen, R. Zhao, J. Li, Y. Gong.
Internal language model estimation for domain-adaptive end-to-end speech recognition.
ArXiv, Vol. abs/2011.01991, 2020.

[Michel & Schlüter+ 20] W. Michel, R. Schlüter, H. Ney.
Early Stage LM Integration Using Local and Global Log-Linear Combination.
In *Proc. Interspeech 2020*, pp. 3605–3609, 2020.

[Rousseau & Deléglise+ 14] A. Rousseau, P. Deléglise, Y. Estève.
Enhancing the TED-LIUM corpus with selected data for language modeling and more TED talks.

References

In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pp. 3935–3939, Reykjavik, Iceland, May 2014. European Language Resources Association (ELRA).

[Variani & Rybach+ 20] E. Variani, D. Rybach, C. Allauzen, M. Riley.
Hybrid autoregressive transducer (HAT).
In *ICASSP*, 2020.