RELEVANT LINKS

DATE
06/11/2025

BLOCKDECODER: BOOSTING ASR DECODERS WITH **CONTEXT** AND **MERGER** MODULES

Darshan Prabhu Preethi Jyothi

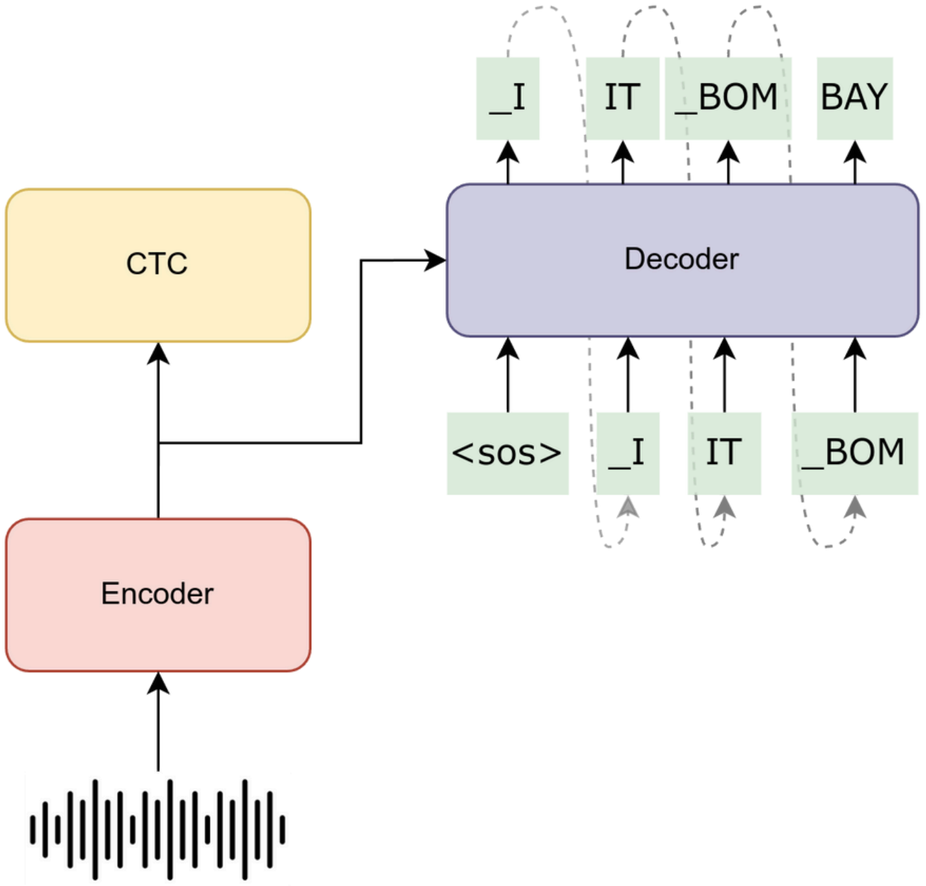
NEURAL INFORMATION PROCESSING SYSTEMS

PRESENTED BY
Darshan Prabhu

RELEVANT LINKS

DATE
06/11/2025

The Architecture of Focus

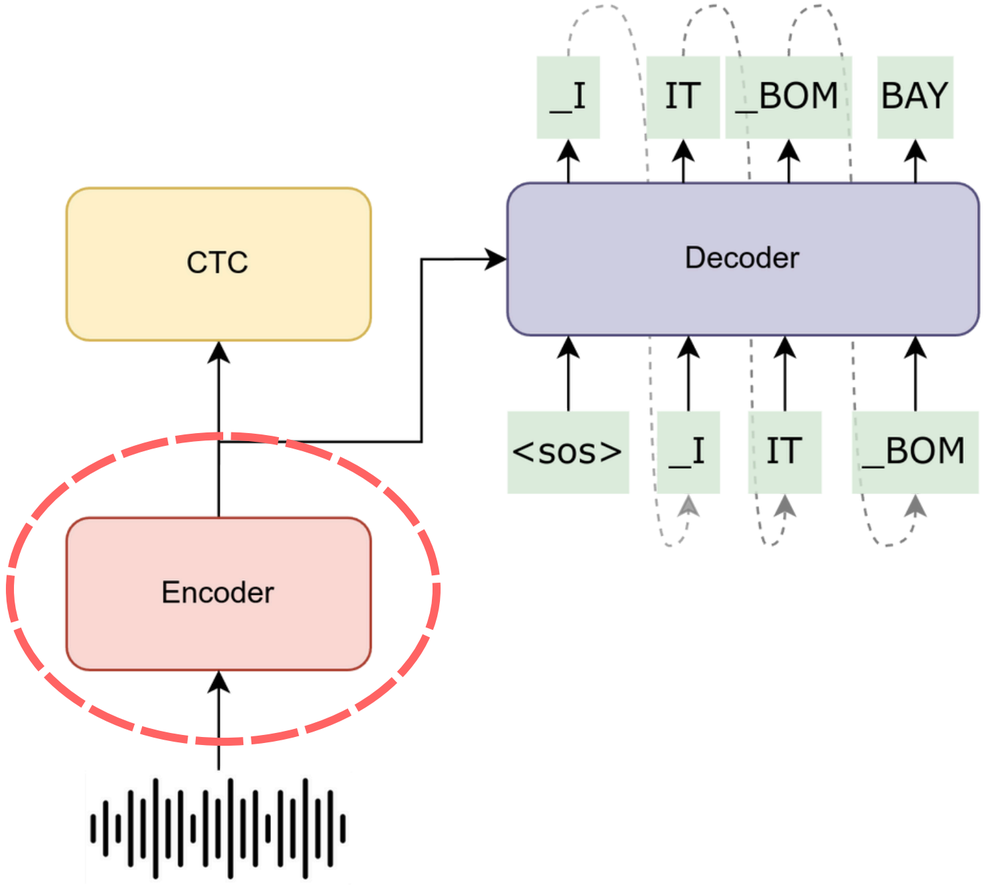


Reference: Watanabe, Shinji & Hori, Takaaki & Kim, Suyoun & Hershey, John & Hayashi, Tomoki. (2017). Hybrid CTC/Attention Architecture for End-to-End Speech Recognition. IEEE Journal of Selected Topics in Signal Processing. PP. 1-1. 10.1109/JSTSP.2017.2763455.

RELEVANT LINKS

DATE
06/11/2025

Prior Literature



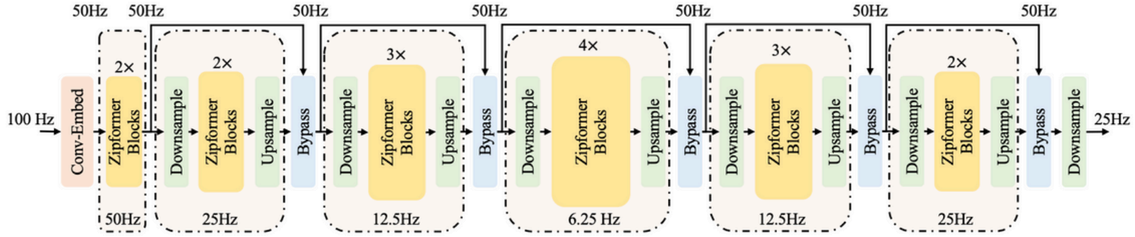
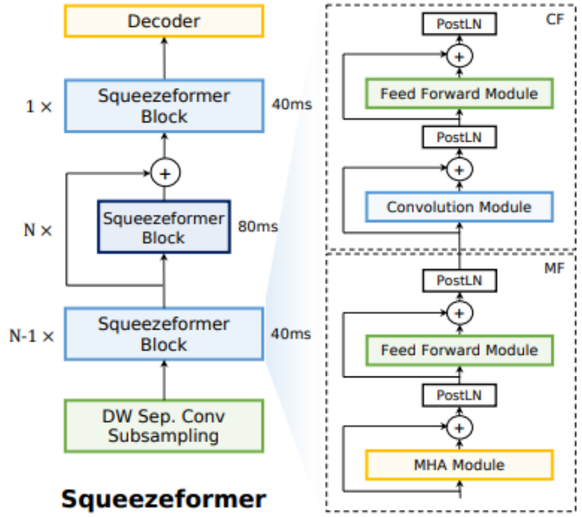


Figure 1: Overall architecture of Zipformer.

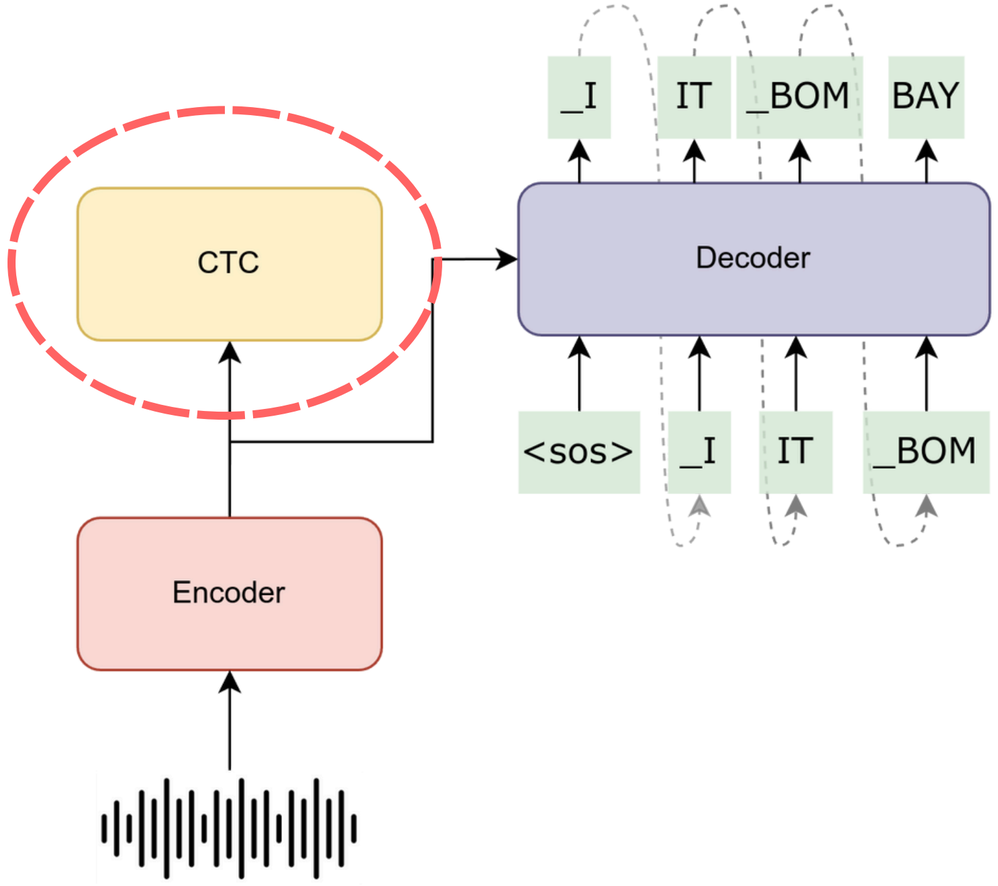


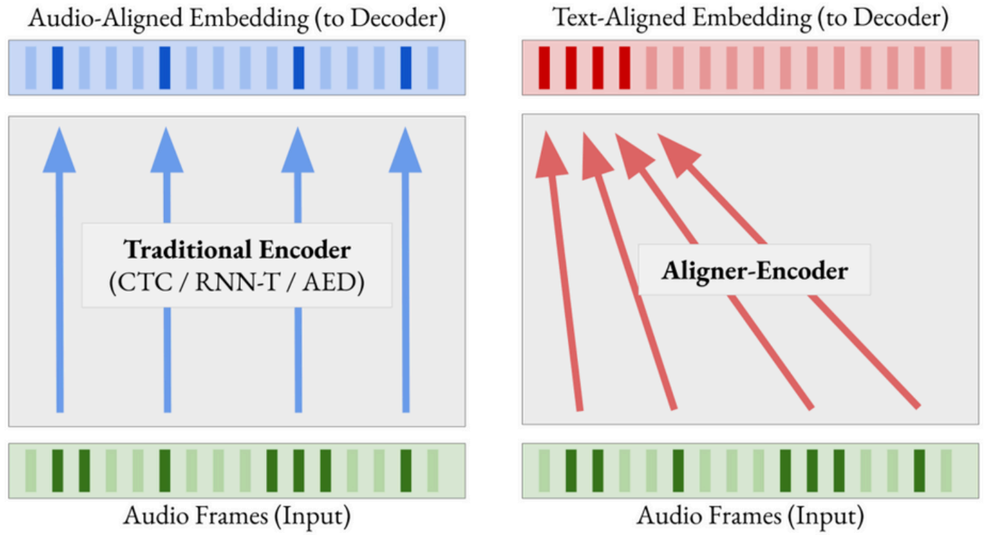
Reference: Yao, Z., Guo, L., Yang, X., Kang, W., Kuang, F., Yang, Y., & Povey, D. (2023). Zipformer: A faster and better encoder for automatic speech recognition. arXiv preprint arXiv:2310.11230. (ICLR 2024)

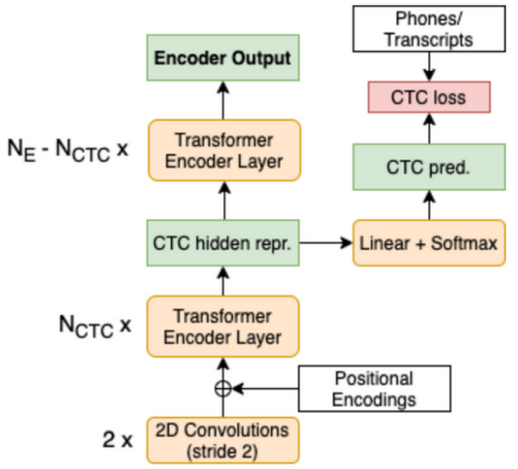
RELEVANT LINKS

DATE
06/11/2025

Prior Literature (contd.)





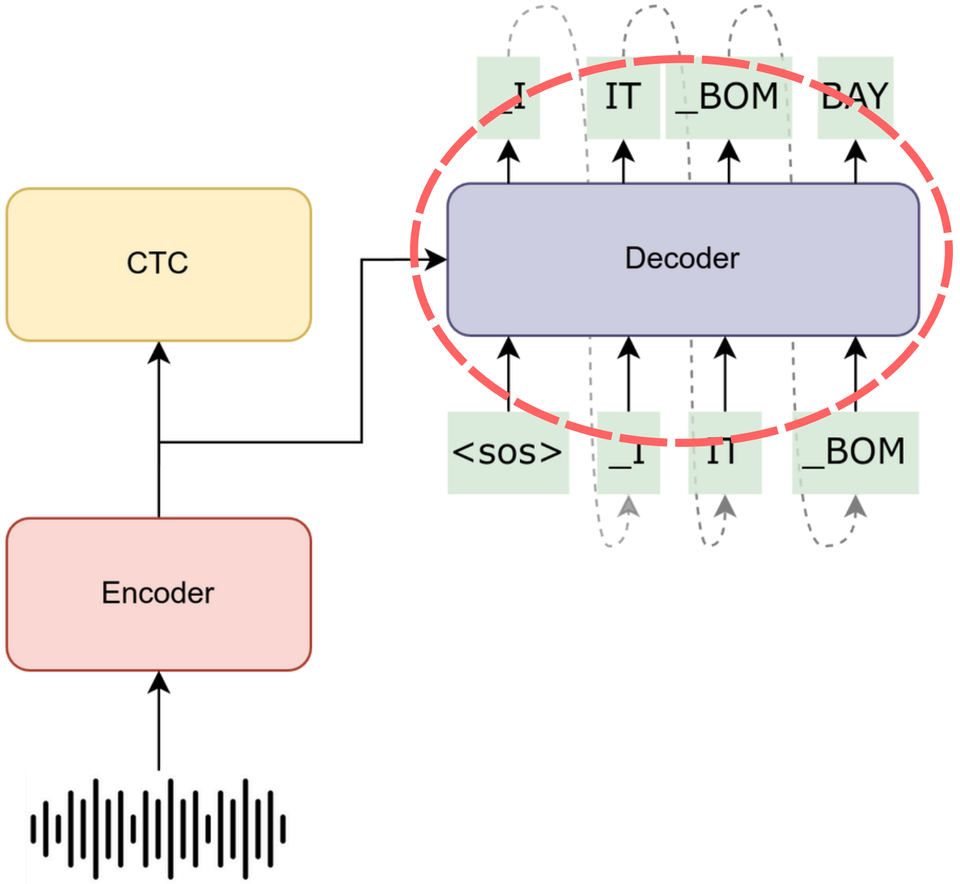


Reference: Stooke, A., Prabhavalkar, R., Sim, K., & Moreno Mengibar, P. (2024). Aligner-Encoders: Self-Attention Transformers Can Be Self-Transducers. Advances in Neural Information Processing Systems, 37, 100318-100340. (NeurIPS 2024)

RELEVANT LINKS

DATE
06/11/2025

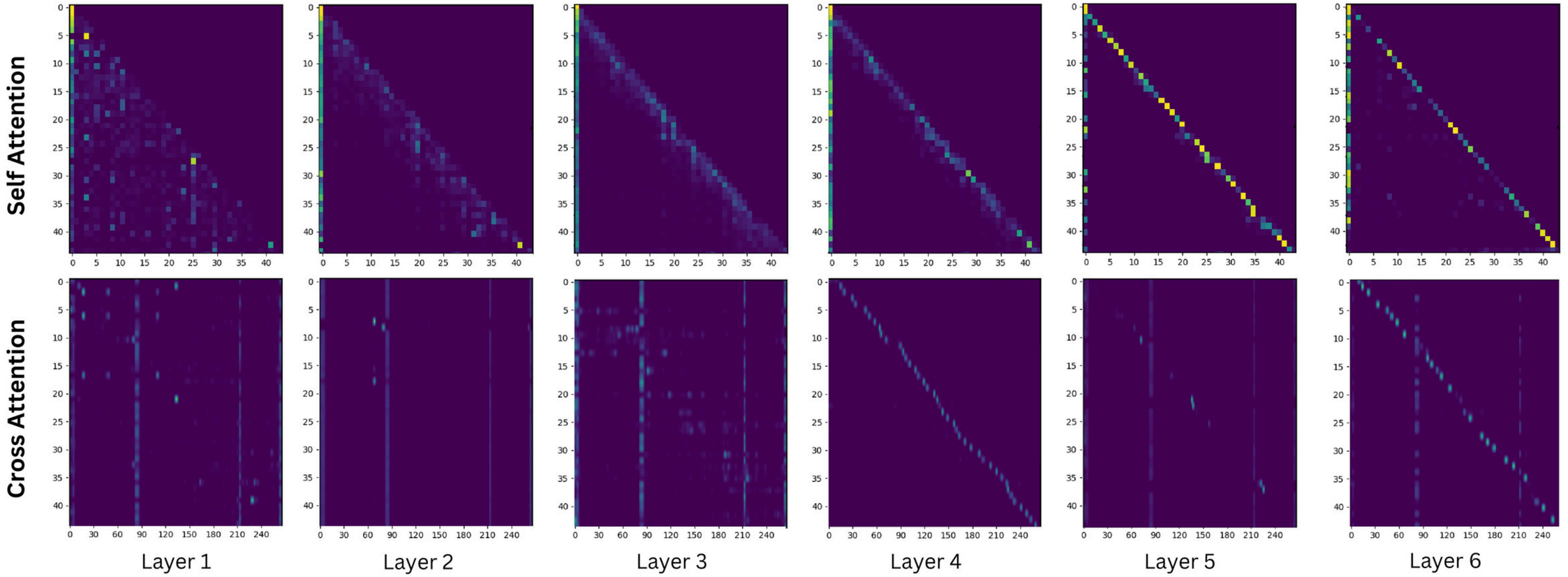
What is missing ?



RELEVANT LINKS

Analysis of attention patterns

DATE
06/11/2025



1 **Self-attention** increasingly focuses on local context as we progress deeper into the decoder.

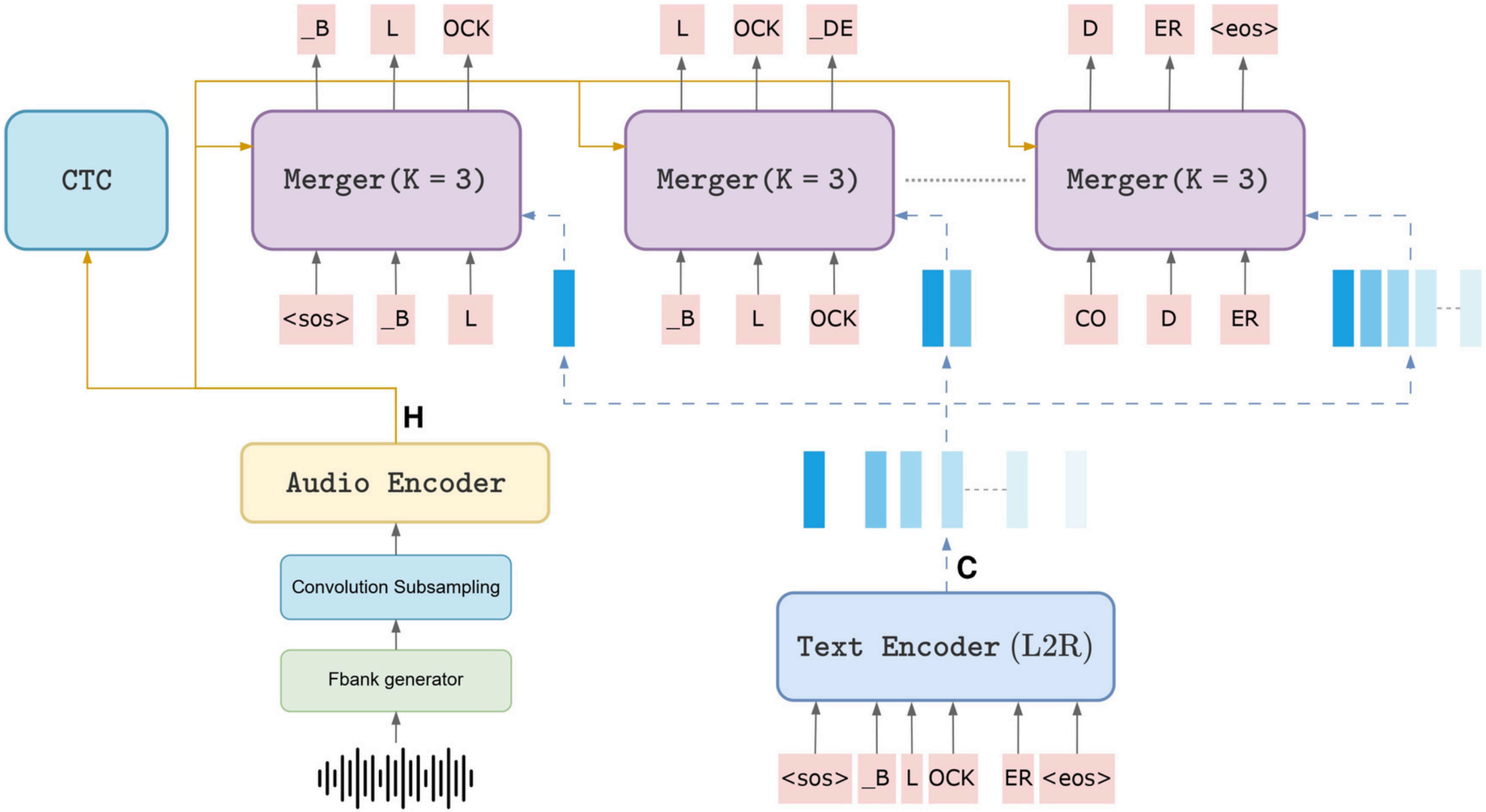
2 **Cross-attention** blocks appear to be less effective in the initial decoder layers.

RELEVANT LINKS

BlockDecoder

Boosting ASR Decoders with Context and Merger Modules

DATE
06/11/2025

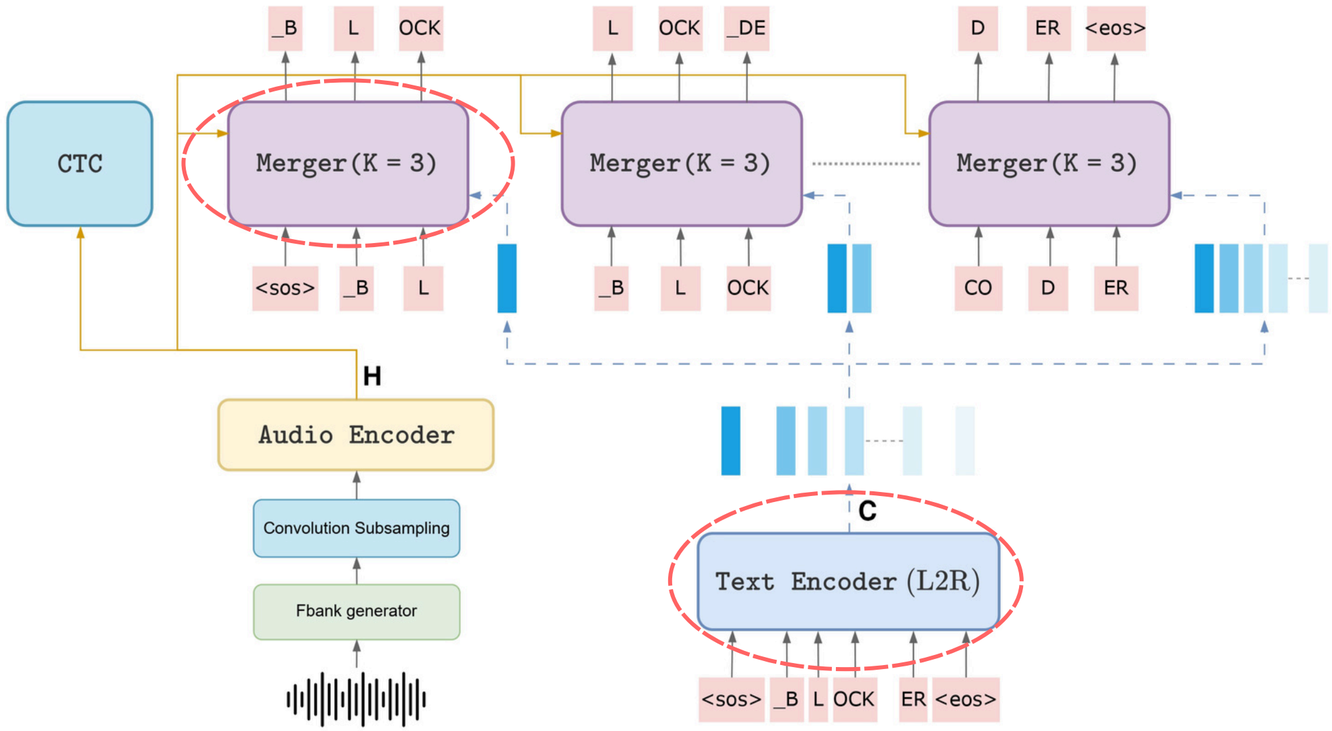


RELEVANT LINKS

BlockDecoder

Boosting ASR Decoders with Context and Merger Modules

DATE
06/11/2025



Key Insights

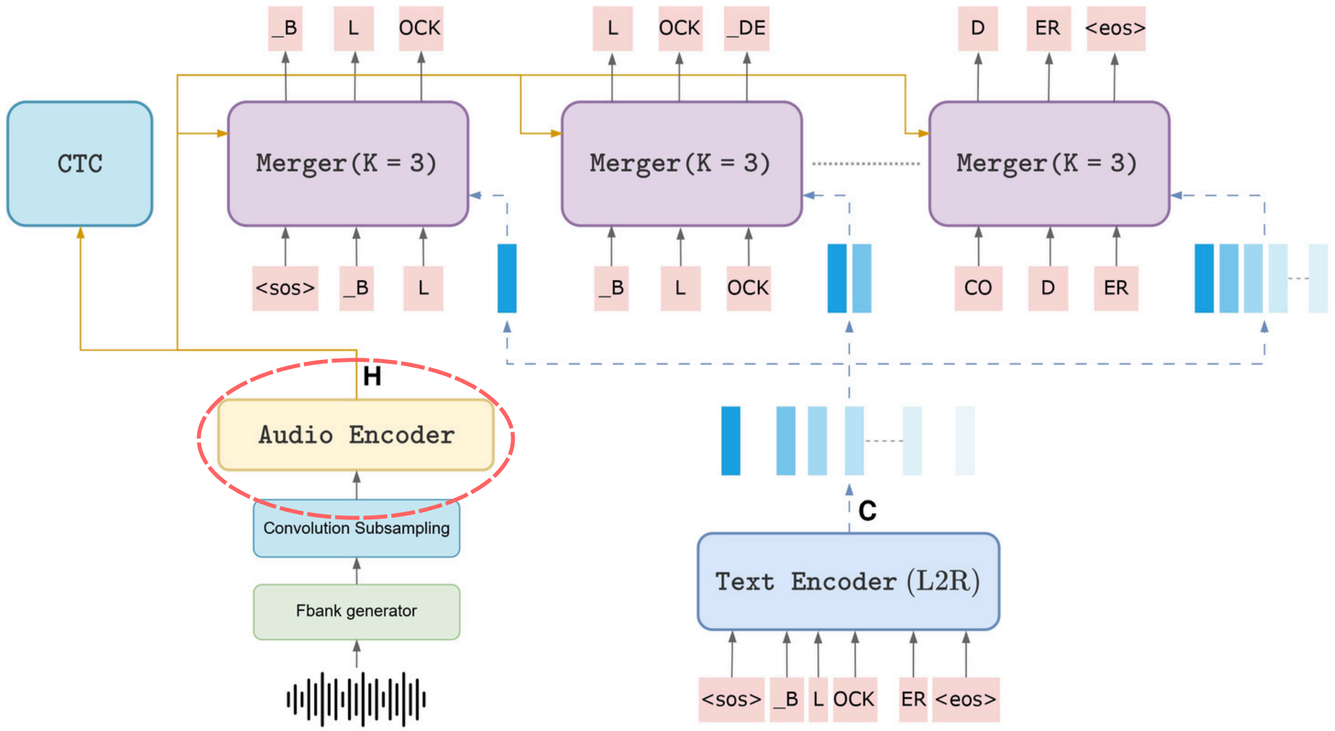
- Replace the Traditional decoder with **Text Encoder** and **Merger**

RELEVANT LINKS

BlockDecoder

Boosting ASR Decoders with Context and Merger Modules

DATE
06/11/2025



Key Insights

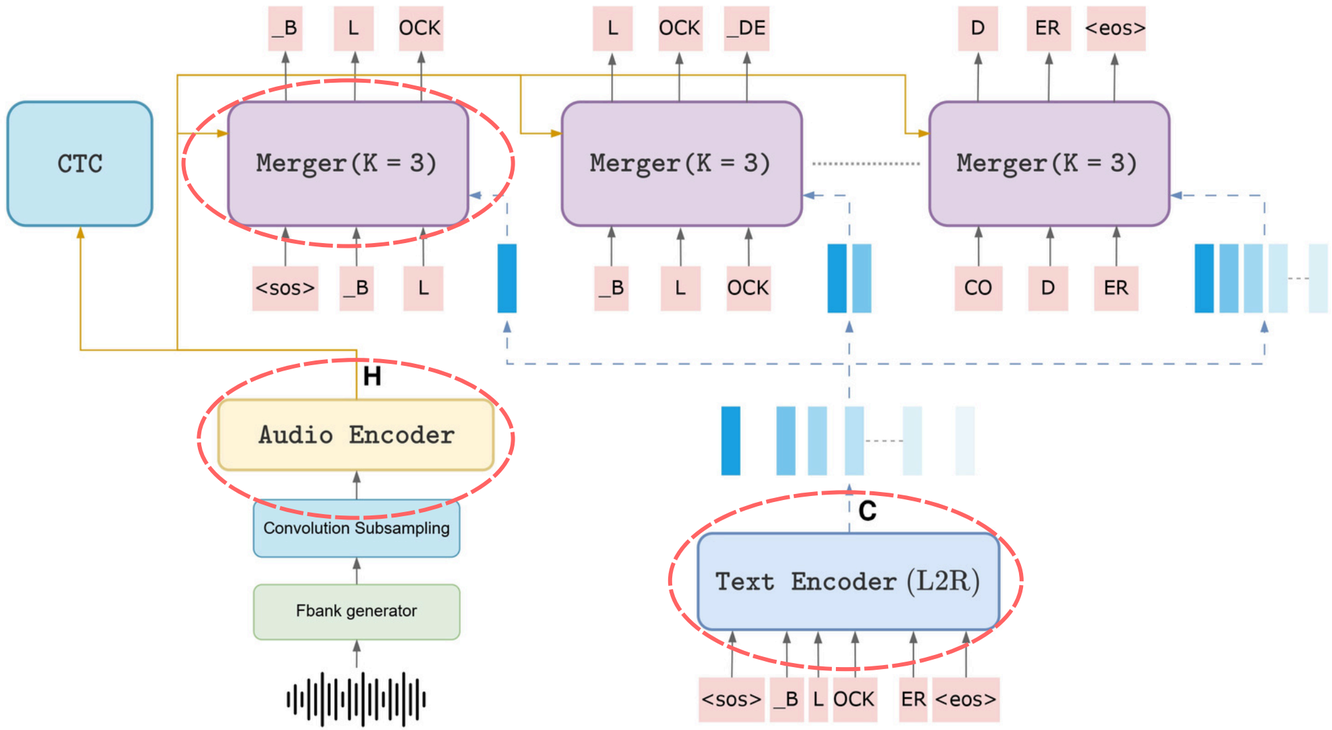
- Replace the Traditional decoder with **Text Encoder** and **Merger**
- **Audio Encoder** → Builds rich audio context

RELEVANT LINKS

BlockDecoder

Boosting ASR Decoders with Context and Merger Modules

DATE
06/11/2025



Key Insights

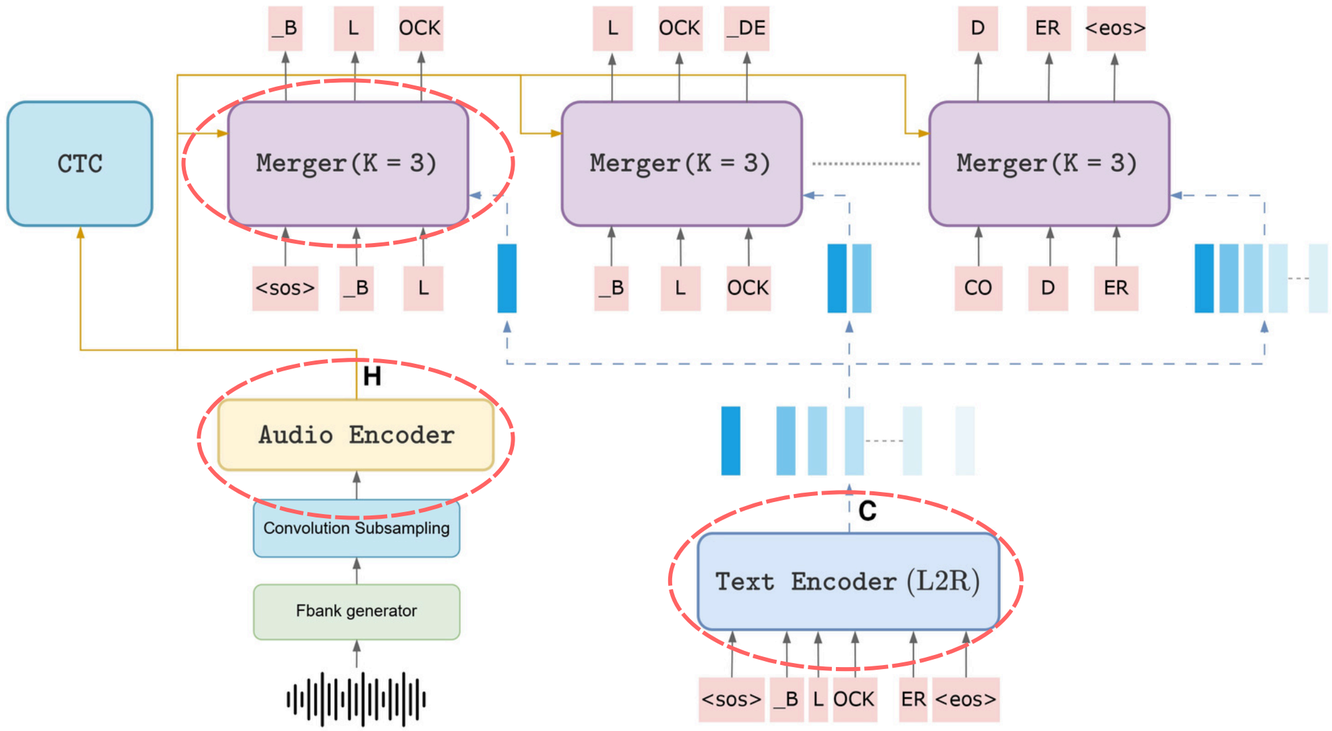
- Replace the Traditional decoder with **Text Encoder** and **Merger**
- **Audio Encoder** → Builds rich audio context
- **Text Encoder** → Builds rich text context, free from cross attention

RELEVANT LINKS

BlockDecoder

Boosting ASR Decoders with Context and Merger Modules

DATE
06/11/2025



Key Insights

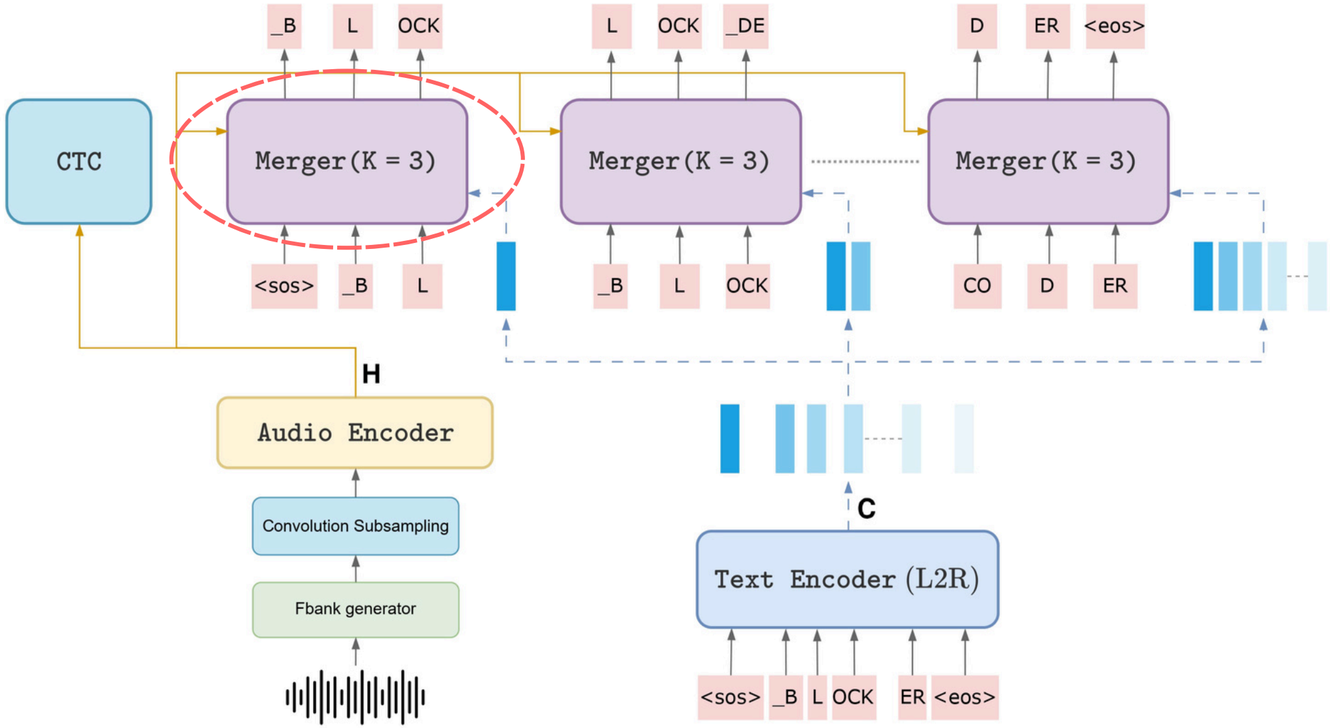
- Replace the Traditional decoder with **Text Encoder** and **Merger**
- **Audio Encoder** → Builds rich audio context
- **Text Encoder** → Builds rich text context, free from cross attention
- **Merger** → Combines contexts and auto-regressively predicts a **block of K tokens**

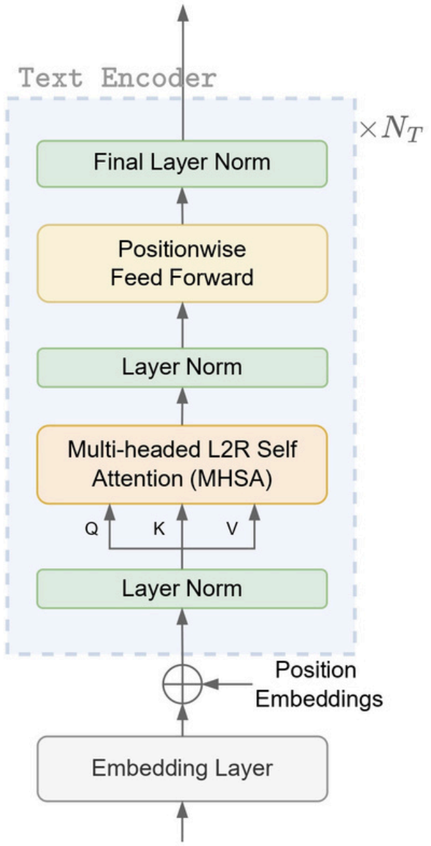
RELEVANT LINKS

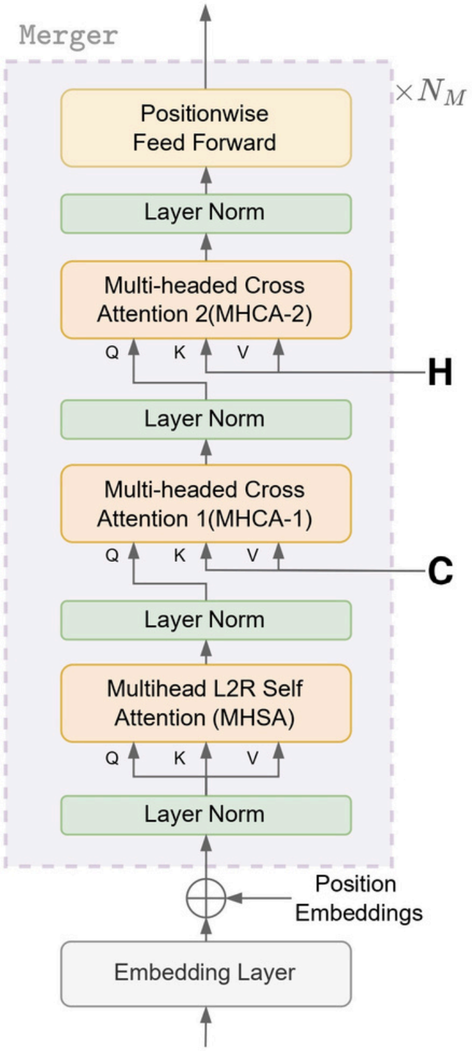
BlockDecoder

Boosting ASR Decoders with Context and Merger Modules

DATE
06/11/2025





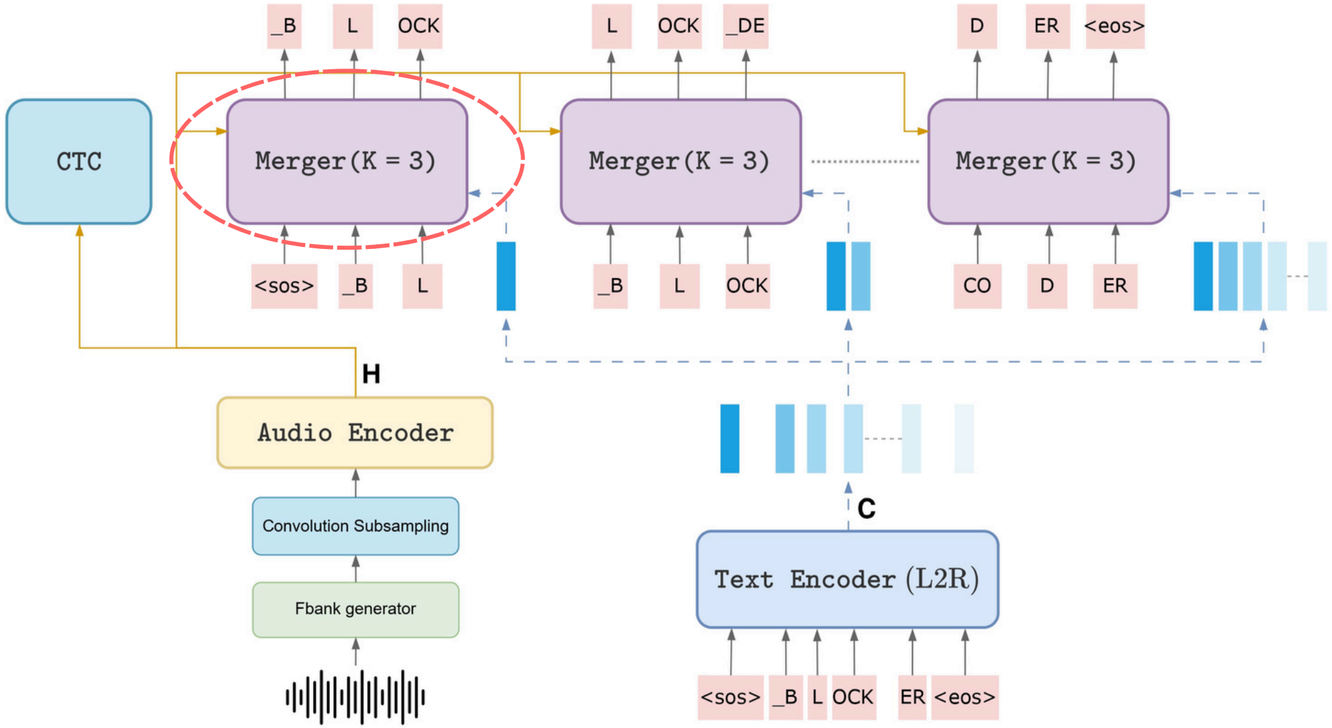


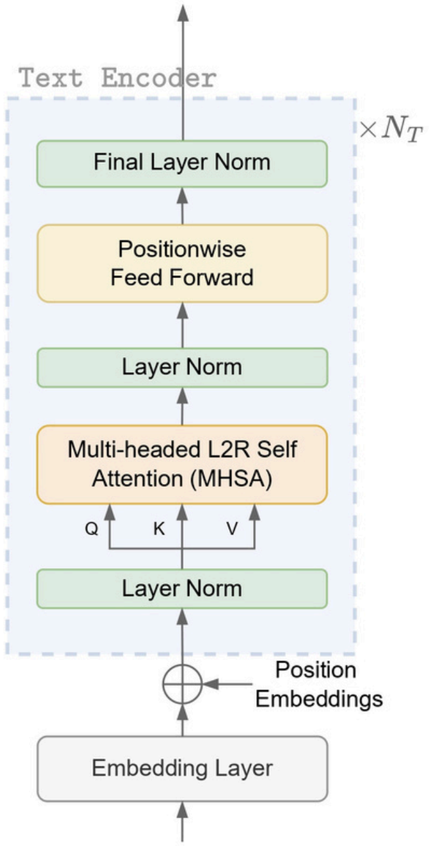
RELEVANT LINKS

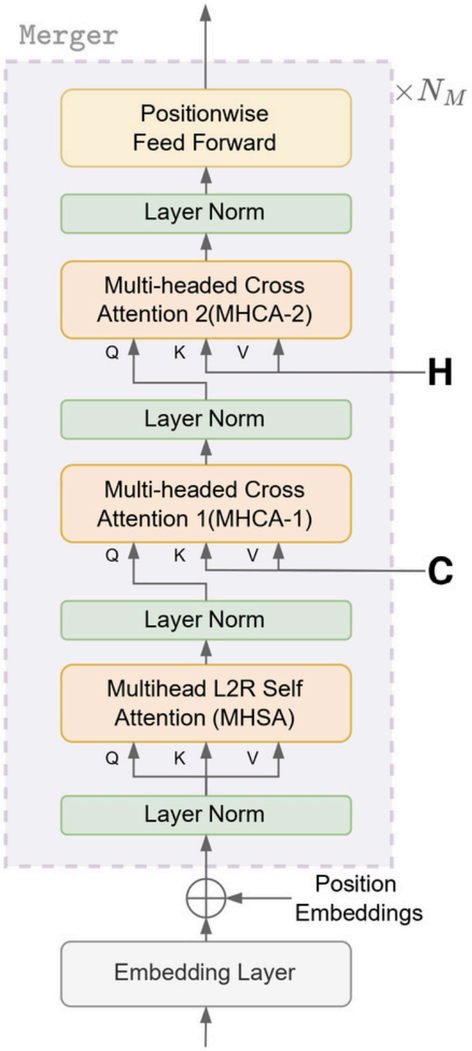
BlockDecoder

Boosting ASR Decoders with Context and Merger Modules

DATE
06/11/2025







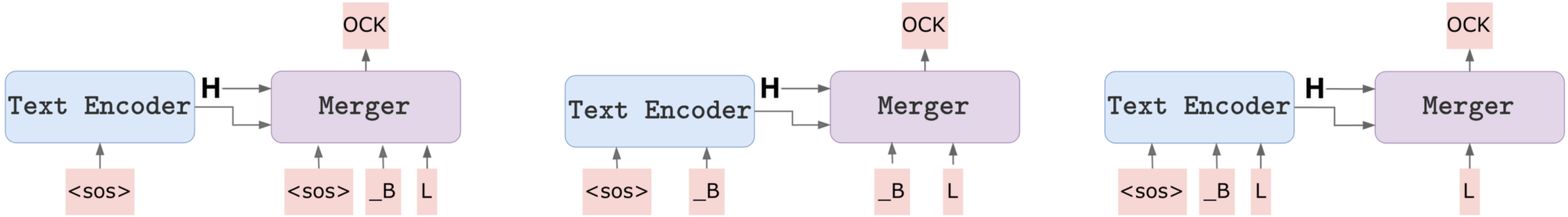
With rich contexts, a very small number of merger layers are enough, thereby reducing parameters.

RELEVANT LINKS

BlockDecoder

Boosting ASR Decoders with Context and Merger Modules

DATE
06/11/2025



(a) (b) (c)

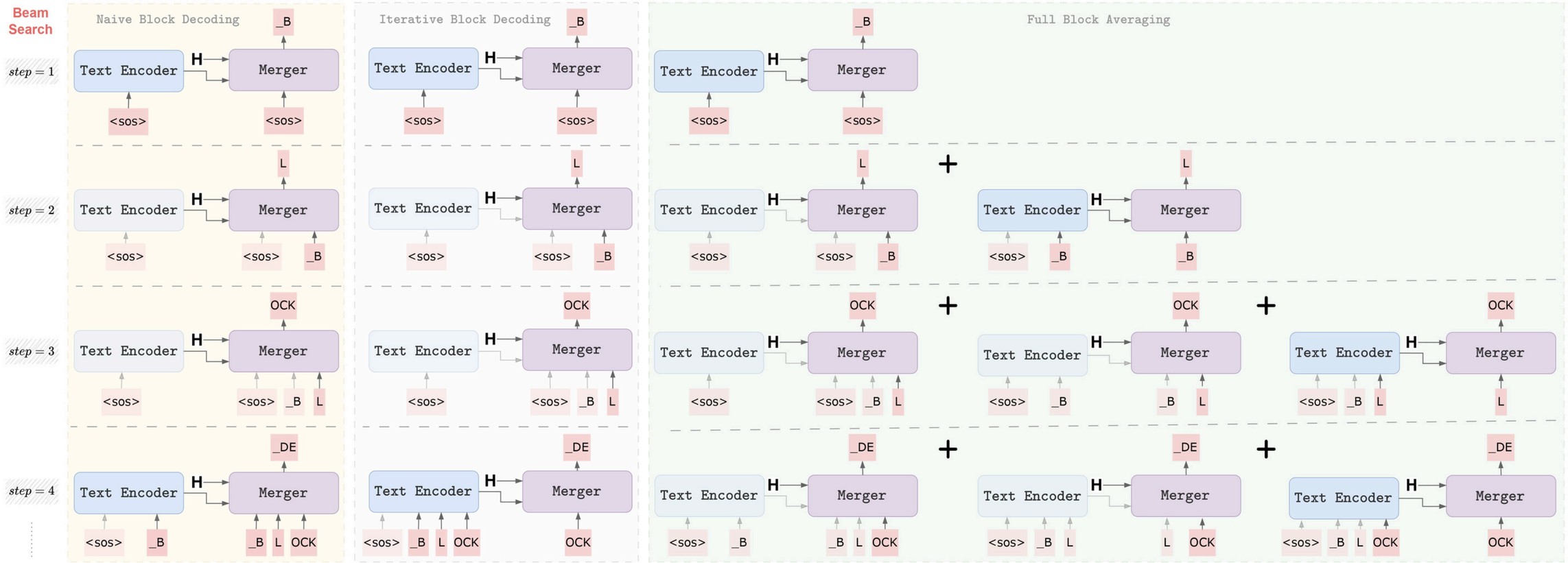
$$P(\mathrm{OCK}/ < \mathrm{SOS} >, \mathrm{\_B}, \mathrm{L})$$

RELEVANT LINKS

BlockDecoder

Boosting ASR Decoders with Context and Merger Modules

DATE
06/11/2025



RELEVANT LINKS

BlockDecoder

Boosting ASR Decoders with Context and Merger Modules

DATE
06/11/2025

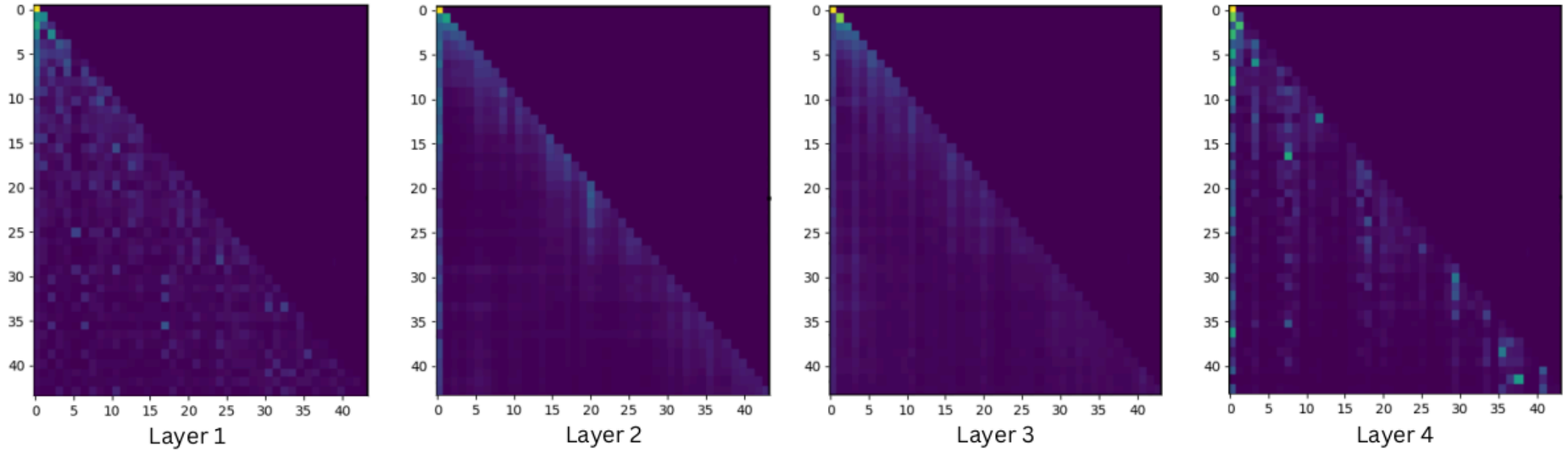
Method	Librispeech-100h (WER)				Tedlium2 (WER)			AISHELL (CER)		
	Params (M)	Test Clean ↓	Test Other ↓	RTF ↓	Params (M)	Test ↓	RTF ↓	Params (M)	Test ↓	RTF ↓
Conformer [10]										
w/ Transformer Decoder	34.2	6.75	17.74	1.38	30.8	**7.69**	2.16	33.6	**4.58**	0.44
w/ BlockDecoder										
– Naive Block Decoding	33.7	6.63	17.63	$0.73_{(\sim 1.9x)}$	30.2	7.81	$1.02_{(\sim 2.1x)}$	33.1	4.76	$0.28_{(\sim 1.6x)}$
– Iterative Block Decoding	33.7	6.72	17.70	$\mathbf{0.66}_{(\sim 2.1x)}$	30.2	7.92	$\mathbf{0.90}_{(\sim 2.4x)}$	33.1	4.75	$\mathbf{0.25}_{(\sim 1.8x)}$
– Full Block Averaging	33.7	**6.58**	**17.62**	1.68 ▷	30.2	7.79	2.39 ▷	33.1	4.63	0.51 ▷
E-Branchformer [21]										
w/ Transformer Decoder	38.5	6.39	17.03	1.52	35.0	**7.44**	2.17	37.9	**4.50**	0.45
w/ BlockDecoder										
– Naive Block Decoding	37.9	6.15	**16.82**	$0.77_{(\sim 2.0x)}$	34.5	7.61	$1.04_{(\sim 2.1x)}$	37.4	4.61	$0.30_{(\sim 1.5x)}$
– Iterative Block Decoding	37.9	6.19	16.94	$\mathbf{0.67}_{(\sim 2.3x)}$	34.5	7.60	$\mathbf{0.91}_{(\sim 2.4x)}$	37.4	4.61	$\mathbf{0.26}_{(\sim 1.7x)}$
– Full Block Averaging	37.9	**6.14**	16.85	1.68 ▷	34.5	7.61	2.40 ▷	37.4	4.53	0.54 ▷

RELEVANT LINKS

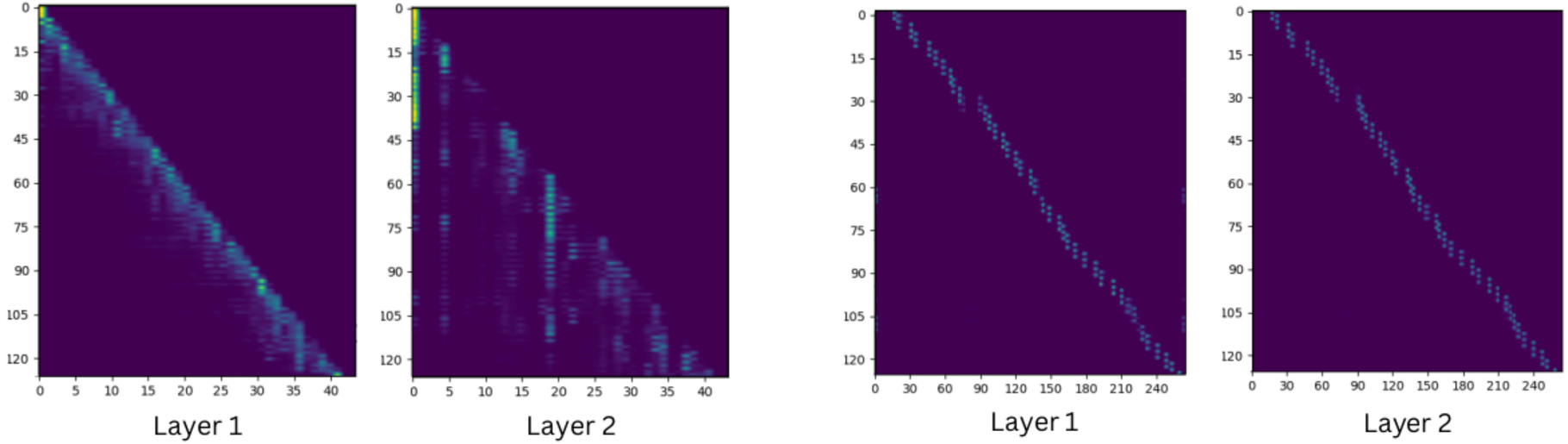
BlockDecoder

Boosting ASR Decoders with Context and Merger Modules

DATE
06/11/2025



(a)



RELEVANT LINKS

Want to know more?

DATE
06/11/2025

https://openreview.net/forum?id=cGmcHJEFnY

Email: darshanp@cse.iitb.ac.in

POSTER SESSION

Exhibit Hall C,D,E

Wed 3 Dec 11 a.m. – 2 p.m. PST

If you are interested, please stop by. 🙂

RELEVANT LINKS

DATE
06/11/2025

https://openreview.net/forum?id=cGmcHJEFnY

Thank you!

Darshan Prabhu Preethi Jyothi

Email: darshanp@cse.iitb.ac.in