Neural Similarity Measurement Strategies in Speaker Diarization Task

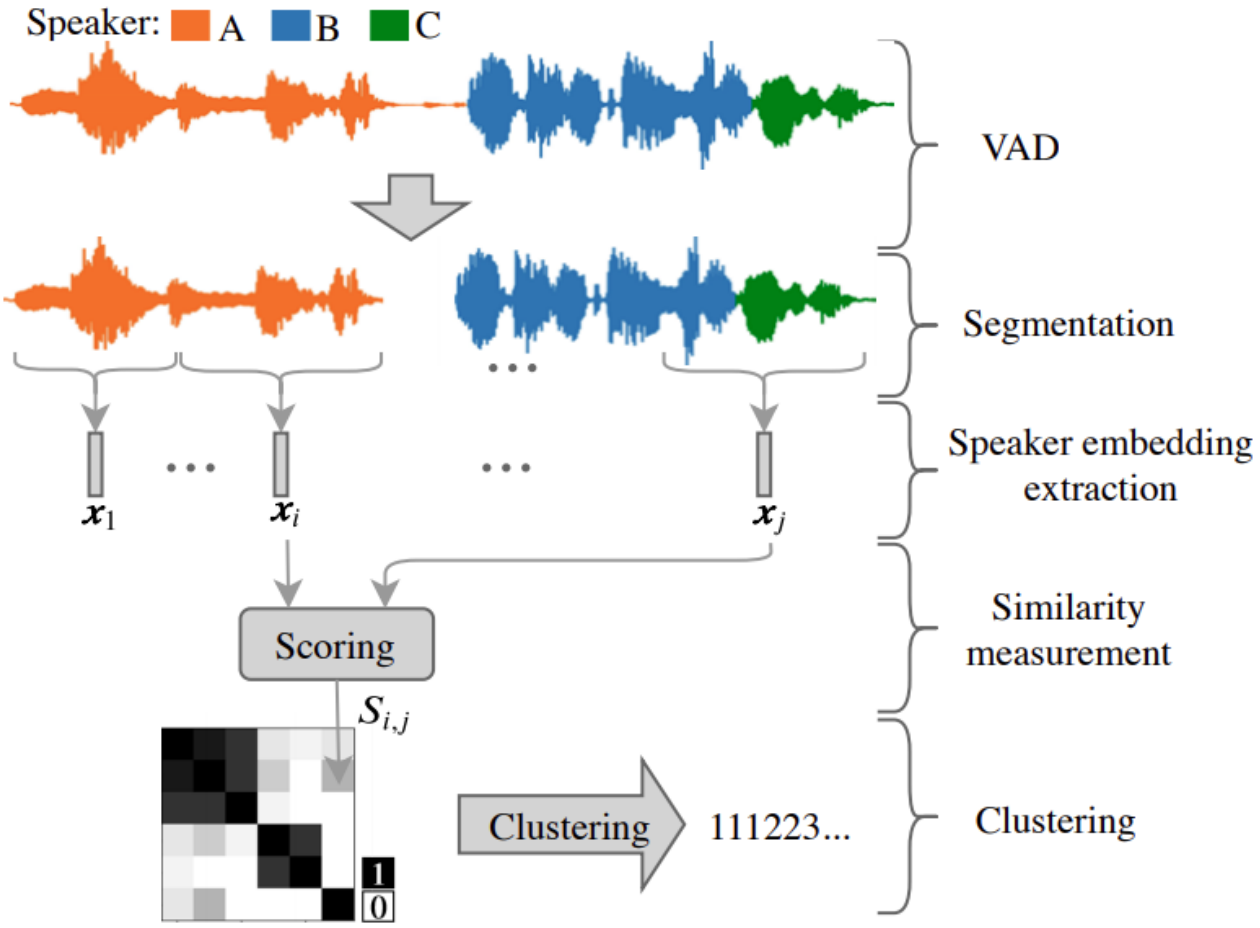
Nikita Kuzmin

Lomonosov Moscow State University

November 17, 2020

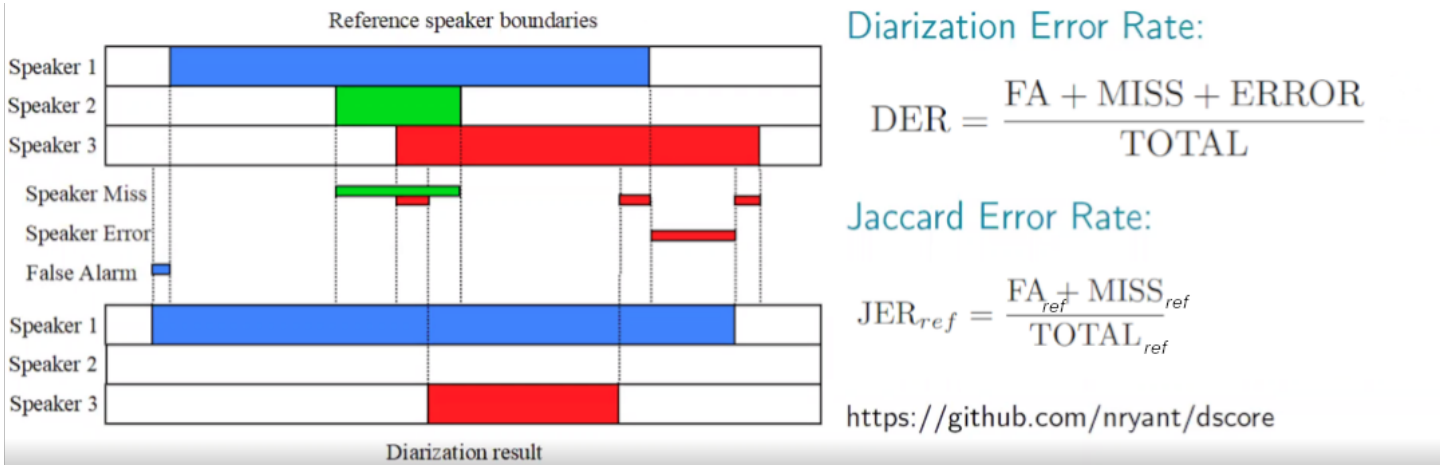
Introduction

Speaker diarization solves "who spoke when" problem. Typical pipeline is presented below:



Metrics

Typical diarization metrics:



$JER = \frac{1}{N} \sum_{ref} JER_{ref}$

- **TOTAL** is the total reference speaker time; that is, the sum of the durations of all reference speaker segments;
- **FA** is the total system speaker time not attributed to a reference speaker;
- **MISS** is the total reference speaker time not attributed to a system speaker;
- **ERROR** is the total reference speaker time attributed to the wrong speaker.

Typical Datasets

- **NIST SRE 2000 CALLHOME** - contains 500 utterances in total, in 6 languages: English, Chinese, Japanese, Arabic, German, and Spanish. The number of speakers in the recordings, ranges from 2 to 7 speakers;
- **DIHARD (dev and eval)** - contains recordings from different domains (child language, clinical, courtroom, etc...);

Input condition	Set	Duration (hours)	# Recordings
single channel	dev	23.81	192
	eval	22.49	194
multichannel	dev	262.41	105
	eval	31.24	12

- **AMI** and **ICSI** (meetings) - contain about 170 hours in total. Audios are recorded in 16k sample rate, and the average duration is around 40 minutes.

Traditional Approaches

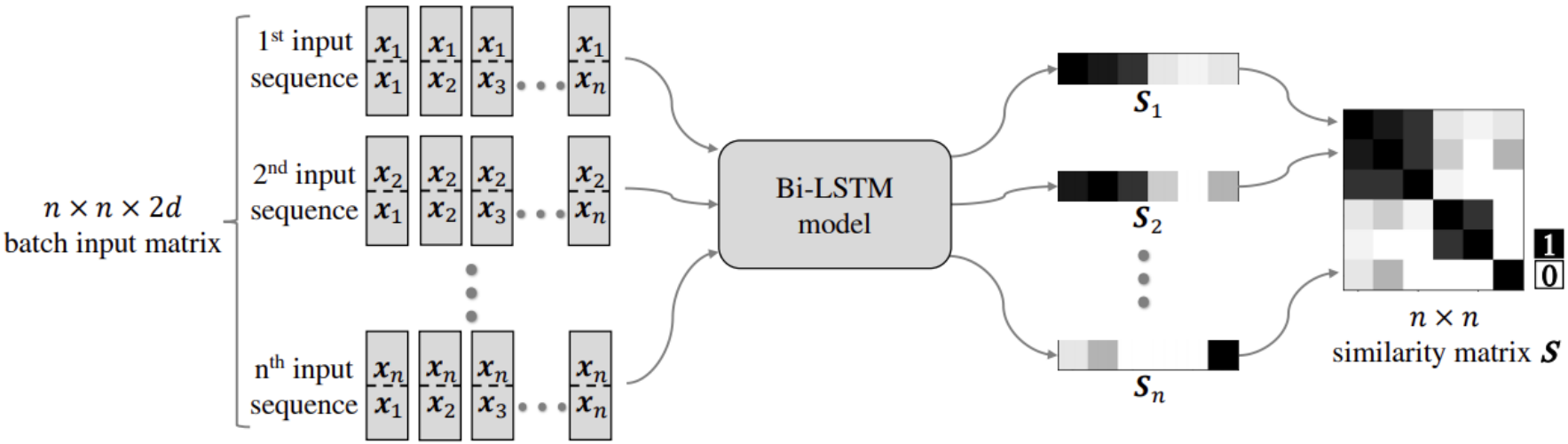
We need to compute similarity matrix $S \in \mathbb{R}^{n \times n}$,
$\{S\}_{ij} = sim(e_i, e_j)$, where $e_i, e_j \in \mathbb{R}^d$ - i-th and j-th embeddings respectively; $sim(\cdot, \cdot)$ - similarity function.
Examples of $sim(x, y)$:

- **Cosine** (works well with metric learning);
- **PLDA**;
- Minkowski.

LSTM based vector-to-sequence scoring

Problem: Cosine, PLDA works in a pair-wise and independent maner $\Rightarrow$ ignore positional correlation.

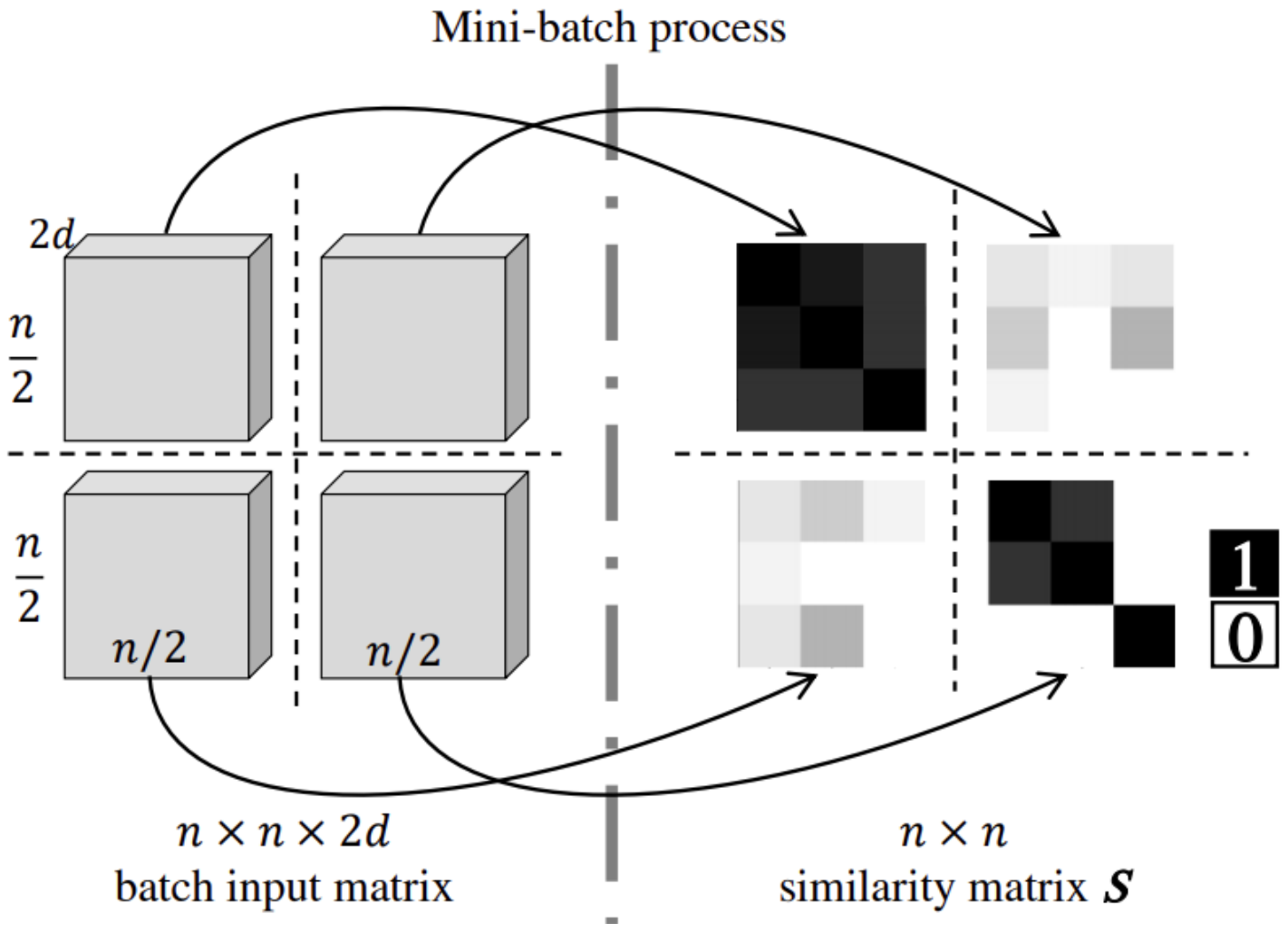
Solution: Use LSTM based vector-to-sequence scoring:



Problems of Solution:

1. If n is too large? Memory problem;
2. LSTM model forgets about previous embeddings;
3. Inference stage is too slow.

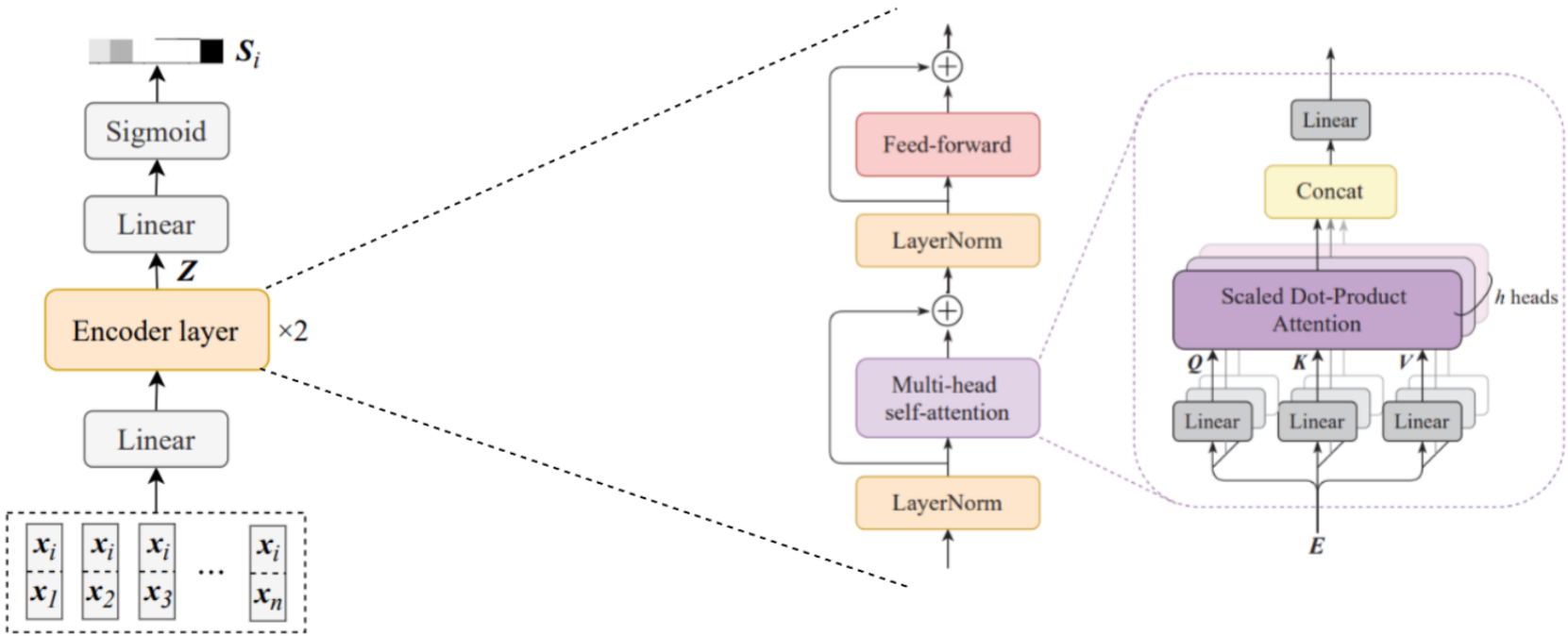
If n is too large?



Self-Attention based vector-to-sequence scoring

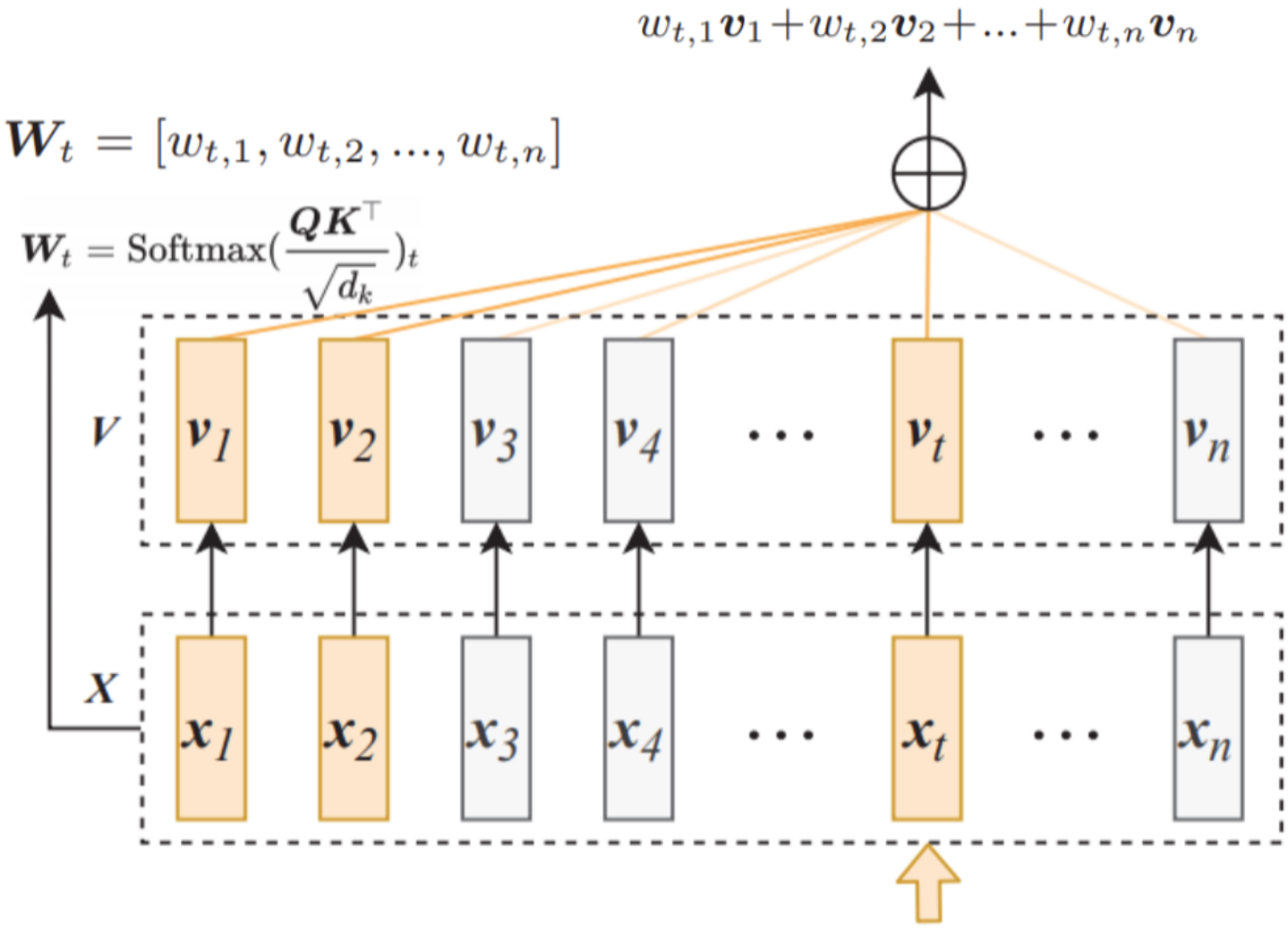
Solution of 2nd LSTM problem

Is something strange on the left picture?



Self-Attention based vector-to-sequence scoring

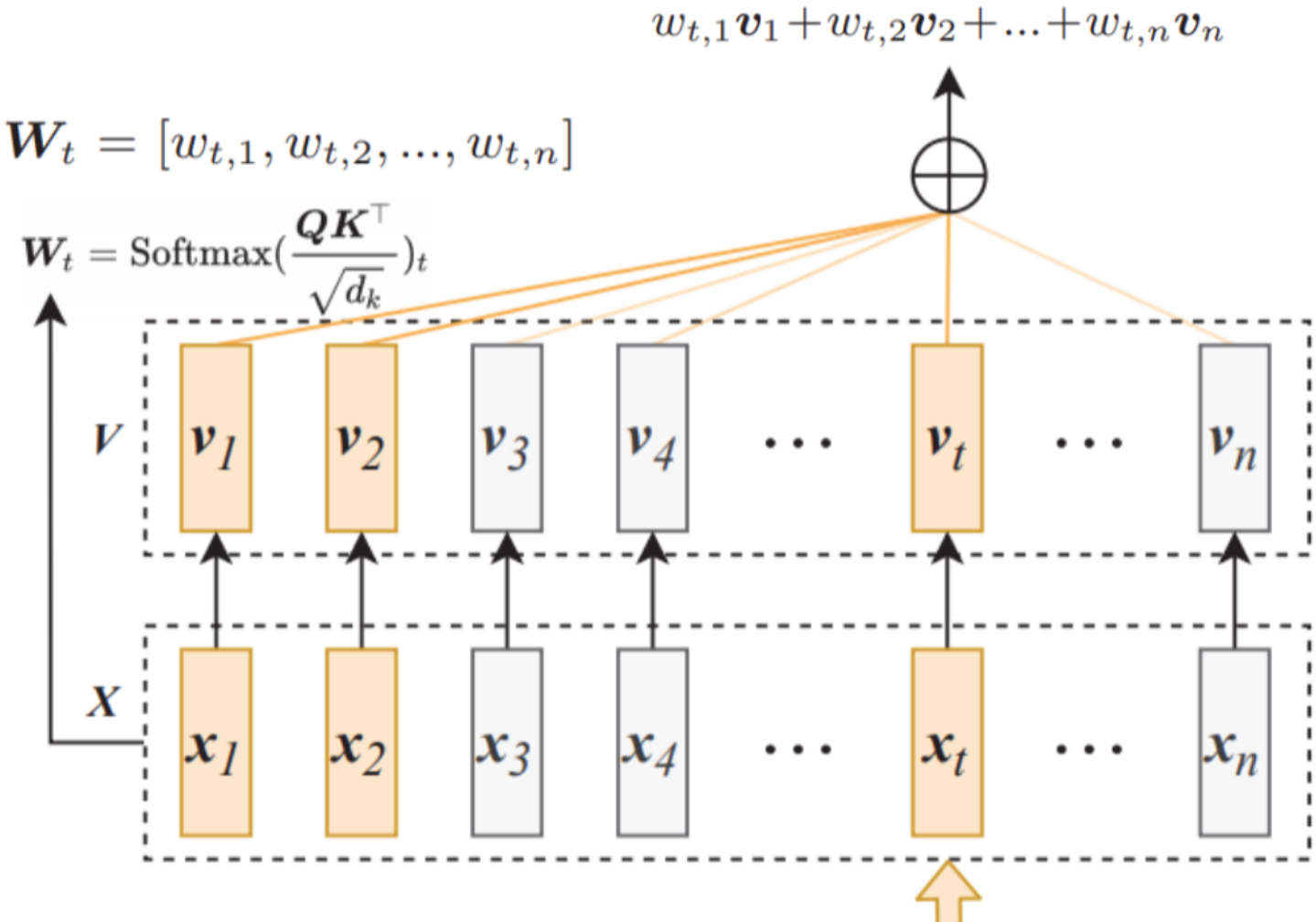
How self-attention mechanism works in speaker diarization:



Self-Attention based vector-to-sequence scoring

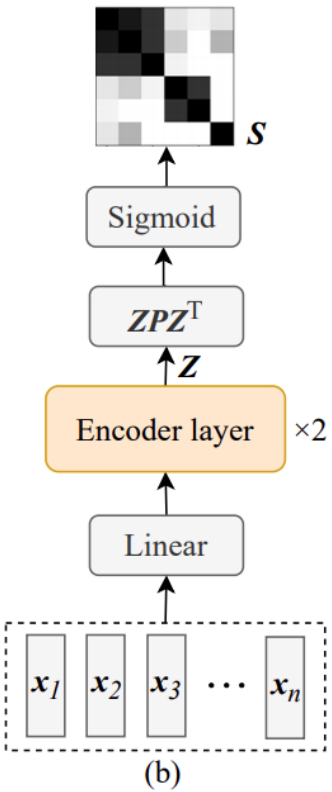
Problem:

- Inference stage even slower, than inference with LSTM.



Self-Attention based sequence-to-sequence scoring

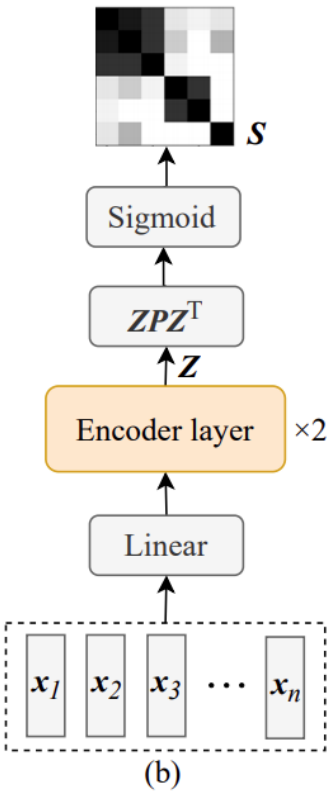
Problem: Self-Attention based vector-to-sequence (v2s) scoring beats SotA in DIHARD II challenge, but inference takes a lot of time;
Authors proposed Self-Attention based sequence-to-sequence (s2s) scoring method



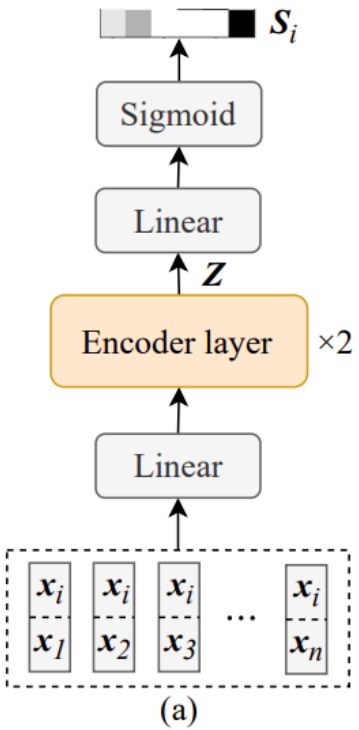
$$ZPZ^\top = \begin{bmatrix} z_1^\top P z_1 & \cdots & z_1^\top P z_n \\ \vdots & \ddots & \vdots \\ z_n^\top P z_1 & \cdots & z_n^\top P z_n \end{bmatrix}$$

Self-Attention based sequence-to-sequence scoring

Problem: Self-Attention based vector-to-sequence (v2s) scoring beats SotA in DIHARD II challenge, but inference takes a lot of time;
Authors proposed Self-Attention based sequence-to-sequence (s2s) scoring method



$$ZPZ^\top = \begin{bmatrix} z_1^\top P z_1 & \cdots & z_1^\top P z_n \\ \vdots & \ddots & \vdots \\ z_n^\top P z_1 & \cdots & z_n^\top P z_n \end{bmatrix}$$



Embeddings

How to get embeddings x_i?

From models trained on Speaker Verification task (for example - from ResNet embedder):

	full-length	**1.5s * 1**	**1.5s * N**
EER(%)	1.51	6.74	1.98

1.5s * N – case with averaging multiple 1.5s speaker embeddings for each full-length utterance.

Results

Datasets:

- For training embedder model: VoxCeleb1, 2 dev;
- For training scoring models: AMI, ICSI – public meeting corpora (+ DIHARD II DEV set);
- For evaluation: DIHARD II DEV and EVAL sets.

Model	+VB	Dev		Eval		Eval + adaptation		Time cost (Eval)
		DER(%)	JER(%)	DER(%)	JER(%)	DER(%)	JER(%)	
LSTM	×	19.65	49.60	20.57	50.25	19.72	46.49	67 min
	√	19.48	49.21	19.98	49.42	19.26	45.91	-
Att-v2s	×	**19.07**	**47.43**	**20.15**	**47.84**	**18.98**	43.20	148 min
	√	**18.76**	**46.77**	**19.46**	**47.01**	**18.44**	42.52	-
Att-s2s	×	19.39	48.42	21.46	48.71	21.45	**43.19**	24 s
	√	19.16	47.99	20.78	47.92	20.12	**41.73**	-
PLDA	×	23.48	57.17	-	-	23.73	56.84	51 s
DIHARD II winner system [27]						18.42	44.58	
DIHARD II official baseline [28]						25.99	59.51	

Results

Thank you for your attention!