TB-RESNET: BRIDGING THE GAP FROM TDNN TO RESNET IN AUTOMATIC SPEAKER VERIFICATION WITH TEMPORAL-BOTTLENECK ENHANCEMENT

Sunmook Choi [1] Sanghyeok Chung [1] Seungeun Lee [1] Soyul Han [2] Taein Kang [2] Jaejin Seo [2] Il-Youp Kwak [2] Seungsang Oh [1]

[1]Department of Mathematics, Korea University [2]Department of Statistics and Data Science, Chung-Ang University

Introduction

Automatic speaker verification (ASV) with VoxCeleb Speaker Recognition dataset

Limitations of previous studies: Employing ResNet-based models for ASV posseses a potential loss of temporal information, thereby casting doubt on its compatibility with statistics pooling methods.

Our Approach: Design an ASV system that **enriches temporal information** resulting in **more meaningful statistics through the statistics pooling layer**.

Methodology

Feature Engineering

- Input: Log-mel spectrograms
- Data Augmentation: MUSAN, Room Impulse Response (RIR), SpecAugment

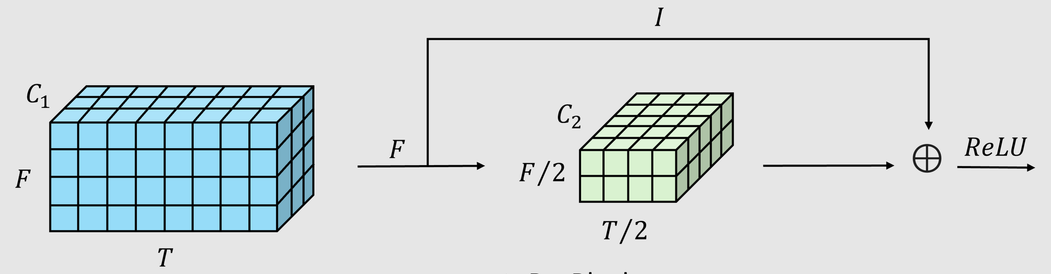
Attentive Statistical Pooling (ASP)

- Attention score
  $\boldsymbol{s}_t = \boldsymbol{W}_2\,\mathrm{ReLU}(\boldsymbol{W}_1\boldsymbol{h}_t + \boldsymbol{b}_1) + \boldsymbol{b}_2, \quad t = 1, \ldots, T.$
  (T: audio length)
- Attention weight
  $\alpha_{t,c} = \frac{\exp(s_{t,c})}{\sum_{i=1}^{T}\exp(s_{i,c})}, \quad c = 1, \ldots, C.$
  (C: dimension of the target feature space)
- Final concatenated vector $[\mu;\sigma]$
  $\mu_c = \sum_{t=1}^{T}\alpha_{t,c}h_{t,c}$
  $\sigma_c = \sqrt{\sum_{t=1}^{T}\alpha_{t,c}h_{t,c}^2 - \mu_c^2}$ (Okabe et al. 2018).

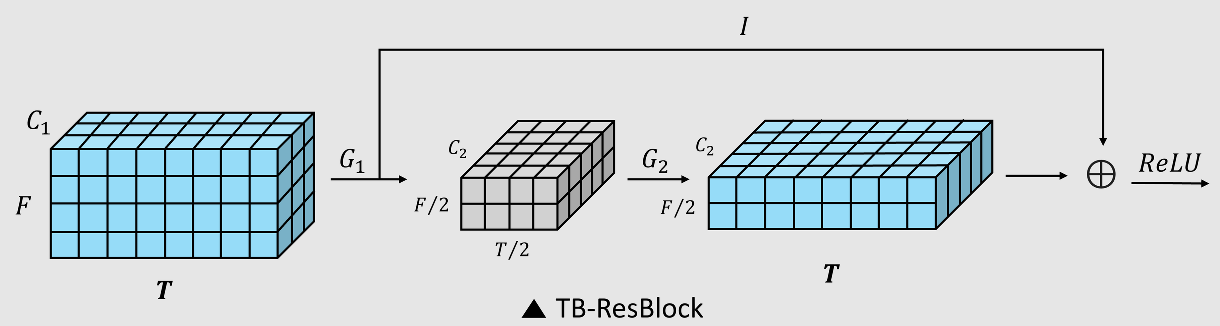
Temporal-Bottleneck Residual Block (TB-ResBlock)

- TB-ResBlock captures and retains temporal information through a transposed convolution by not reducing the corresponding dimension with the series of ResBlocks.

ResBlock vs. TB-ResBlock



▲ ResBlock



▲ TB-ResBlock

- F : 3×3 conv, BN, ReLU, 3×3 conv, BN.
- G_1 : 3×3 conv (stride: $(s,2)$), BN, ReLU.
- G_2 : 3×3 conv (stride: $(1,2)$), BN, ReLU.
- I : 1×1 conv, BN (Identity).
- $s = 1$ or 2. Here, $s = 2$ for readability.

Advantages of TB-ResBlock

- Prevents the reduction of the temporal dimension through TB-ResBlocks → Facilitates effective aggregation in ASP.
- Reduces and subsequently recovers the number of temporal frames → Enables the exploration of more valuable temporal information. (Similar effects observed in bottleneck blocks in deeper ResNet architectures. (He et al. 2016))

Experiment

Experimental Setup

- Training dataset: VoxCeleb2 development set
- Testing dataset: VoxCeleb1 test set (VoxCeleb1-O, VoxCeleb1-H, VoxCeleb1-E) (Nagrani et al. 2019)
- Metric: equal error rate (EER), the minimum detection cost function (minDCF)

Performance Results on VoxCeleb1 test set

Quantitative comparison between ResNet-GAP (global average pooling), ResNet-ASP, and TB-ResNet is shown in **Table 1** below.

Acknowledgement

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Ministry of Science and ICT (RS-2023-00208284) and Institute for Information communications Technology Planning Evaluation (IITP) grant funded by the Korea government(MSIT) (No.2019-0-00033, 50%, Study on Quantum Security Evaluation of Cryptography based on Computational Quantum Complexity).

References

- Nagrani, A., J. S. Chung, W. Xie, and A. Zisserman (2019). "Voxceleb: Large-scale speaker verification in the wild". In: *Computer Science and Language*.
- Okabe, K., T. Koshinaka, and K. Shinoda (2018). "Attentive Statistics Pooling for Deep Speaker Embedding". In: *Proc. Interspeech 2018*, pp. 2252–2256. DOI: 10.21437/Interspeech.2018-993.
- He, K., X. Zhang, S. Ren, and J. Sun (2016). "Deep residual learning for image recognition". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778.

Discussion

Ablation studies

Model	# Frames	VoxCeleb1-O EER (%)	VoxCeleb1-O minDCF
TB-ResNet18	$T/16$	1.68	0.1100
	$T/8$	1.57	0.1050
	$T/4$	1.45	0.0920
	$T/2$	**1.36**	**0.0870**
Bilinear	$T/2$	1.89	0.1155
TB-ResNet34	$T/16$	1.35	0.0850
	$T/8$	1.23	0.0820
	$T/4$	1.27	0.0780
	$T/2$	**1.13**	**0.0687**
Bilinear	$T/2$	1.45	0.0885

Controlling the number of retained temporal frames prior to ASP

- Model tends to achieve better performance when retaining more temporal frames prior to ASP.

Validating the significance of transposed convolution in TB-ResBlocks

- It marks degradation in performance upon the application of bilinear interpolation when compared with that of transposed convolution.

Limitations of TB-ResBlock

- The number of parameters remains unchanged, while there is an increase in computational workload. We will further improve these through additional research.

Tables

Table 1. Model performance

Model	# Params	VoxCeleb1-O EER (%)	VoxCeleb1-O minDCF	VoxCeleb1-H EER (%)	VoxCeleb1-H minDCF	VoxCeleb1-E EER (%)	VoxCeleb1-E minDCF
ResNet18-GAP	11.27M	2.03	0.1410	3.73	0.2300	2.08	0.1410
ResNet18-ASP	13.80M	1.62	0.1109	3.02	0.1842	1.64	0.1100
TB-ResNet18	11.44M	**1.36**	**0.0870**	**2.47**	**0.1497**	**1.39**	**0.0895**
ResNet34-GAP	21.38M	1.60	0.1080	3.20	0.1940	1.73	0.1190
ResNet34-ASP	23.91M	1.35	0.0847	2.69	0.1629	1.43	0.0966
TB-ResNet34	21.55M	**1.13**	**0.0687**	**2.20**	**0.1298**	**1.21**	**0.0779**

Table 2. ResNet Architecture

Layer Name	Layer Details	Output Size
input	-	$(80, T, 1)$
conv1	5×5, BN, ReLU	$(80, T, 64)$
maxpool	3×3 window, stride 2	$(40, T/2, 64)$
conv2_x	Block(64, 1, n_2)	$(40, T/2, 64)$
Box (A / B / C)		
linear	speaker embedding	192

Box A: ResNet with GAP

conv3_x	Block(128, 2, n_3)	$(20, T/4, 128)$
conv4_x	Block(256, 2, n_4)	$(10, T/8, 256)$
conv5_x	Block(512, 2, n_5)	$(5, T/16, 512)$
GAP	-	$(1, 1, 512)$

Box B: ResNet with ASP

conv3_x	Block(128, 2, n_3)	$(20, T/4, 128)$
conv4_x	Block(256, 2, n_4)	$(10, T/8, 256)$
conv5_x	Block(512, 2, n_5)	$(5, T/16, 512)$
flatten	except for time axis	$(T/16, 5\times512)$
ASP	channel-dependent	5120

Box C: TB-ResNet

conv3_x	TB-Block(128, 2, n_3)	$(20, T/2, 128)$
conv4_x	TB-Block(256, 2, n_4)	$(10, T/2, 256)$
conv5_x	TB-Block(512, 2, n_5)	$(5, T/2, 512)$
dw_conv6	5×1, BN, ReLU	$(1, T/2, 512)$
ASP	channel-dependent	1024

felixchoi@korea.ac.kr, ikwak2@cau.ac.kr, seungsang@korea.ac.kr