The bilibili system for VoxCeleb Speaker Recognition Challenge 2023

Xingui Zeng, Zhuo Yang, Shiyi Wan, Wei Deng, XiangCao

bilibili

2023-08-20

Outlines

Track1 System Description

Data preparation

Training Data

Only VoxCeleb2 dev dataset (1,092,009 utterances and 5,994 speakers)

Augmentation

- Offline speaker augmentation:
  - 5-fold speed augmentation based on the Sox speed function (0.8, 0.9,1.0,1.1,1.2; 29970 speakers total)
- Online augmentation: Chain-like augmentation with a probability of 0.6
  - Noise addition augment with MUSAN dataset.
  - RIR reverberation with RIRs dataset
  - Gain augment

Development Data

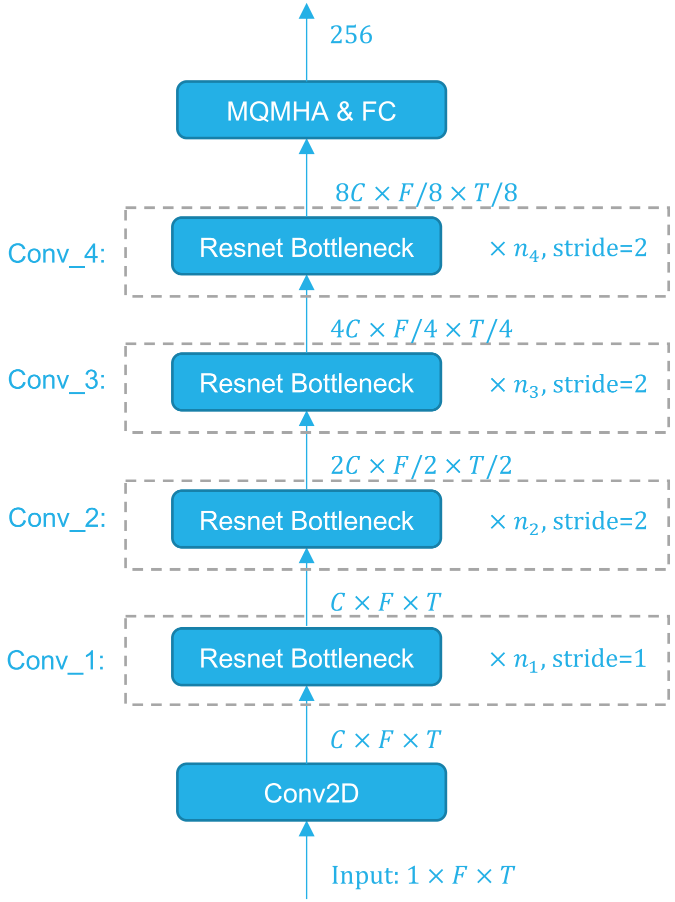
- VoxCeleb1-O
- VoxCeleb1-E
- VoxCeleb1-H
- VoxSRC23-val

Features

- Fbank with {80, 96, 120} dimensions
- w/o additional voice activation detection (VAD)
- w/ cepstral mean normalization(CMN)

Architectures

Backbone



Resnet Variants

Table 1: *Resnet variant*

Name	Features	Resnet Channels	Resnet Depth
$R1$	fbank96	32	$3 \times 8 \times 36 \times 3$
$R2$	fbank120	32	$3 \times 8 \times 36 \times 3$
$R3$	fbank120	64	$3 \times 8 \times 36 \times 3$
$R4$	fbank80	32	$10 \times 20 \times 64 \times 3$

- Four Resnet variants are used.
- To increase the diversity of the models, we make small architectural changes, as shown in Table 1.

Training

Stage 1: Pre-training

- Segment length: 2s
- Optimizer: SGD(momentum 0.9, weight decay 1E-4)
- LR Scheduler: Exponential decrease with warmup(initial lr 0.2, final lr 5r-5)
- Training Objective:
  - AAMSoftmax loss with subcenters and inter-topK penalties
  - gradually increased the margin from 0 to 0.2, scale=32
  - Subcenter: number=3, inter-topK: neighbor=5, penalty=0.06

Stage 2: Large Margin based finetuning

- Segment length: 6s
- Optimizer: SGD(momentum 0.9, weight decay)
- LR Scheduler: Exponential decrease with warmup(initial lr 1e-4, final lr 2.5r-5)
- Training Objective:
  - AAMSoftmax loss with subcenters
  - Margin=0.5, scale=32
  - Subcenter: number=3
- removing the speaker augmentation

Evaluation & Result

Evaluation

- Cosine similarity score was used
- **AS-Norm**: speaker-wise, VoxCeleb2 dev cohorts, top_n=300
- **QMF**:
  - Quality measures: Speech duration, Cosine similarity, AS-normed score, Embedding magnitude
  - we trained an XGBoost to serve as our QMF model.
- **Fusion**: We finetuned the fusion weights of all models based on the results of Voxceleb1-H and VoxSRC 23-val.

Evaluation & Result

- **Ablation study on back-end processing method**

Methods	**EER**	**MinDCF$_{0.05}$**
$R1$	3.221%	0.182
+Large Magin Fintuning	3.073%	0.162
++AS-Norm	2.753%	0.151
+ + +QMF	2.387%	0.141

- **System result**

System	**Voxceleb1-O**		**Voxceleb1-E**		**Voxceleb1-H**		**Voxsrc23-val**		**Voxsrc23-test**	
	EER	**MinDCF$_{0.05}$**	**EER**	**MinDCF$_{0.05}$**	**EER**	**MinDCF$_{0.05}$**	**EER**	**MinDCF$_{0.05}$**	**EER**	**MinDCF$_{0.05}$**
$R1$	0.287%	0.014	0.473%	0.028	0.843%	0.048	2.387%	0.141	2.263%	0.1364
$R2$	0.25%	0.021	0.461%	0.025	0.724%	0.038	2.238%	0.123	-	-
$R3$	0.261%	0.021	0.535%	0.032	0.898%	0.049	2.436%	0.136	-	-
$R4$	0.261%	0.016	0.487%	0.027	0.794%	0.04	2.141%	0.122	-	-
Fusion										
$R1 \sim R4$	0.165%	0.012	0.412%	0.022	0.66%	0.034	1.835%	0.107	1.7810 %	0.1048

Thank you