The SJTU X-LANCE Lab System for CNSRC 2022

X-LANCE Lab, Shanghai Jiao Tong University

Zhengyang Chen, Bei Liu, Bing Han, Leying Zhang, Yanmin Qian

Results

- Speaker Verification Task
  - 0.2975 minDCF, 4.911 EER
- Speaker Retrieval Task
  - 0.4626 mAP

Data Usage

- Training Dataset
  - CN-Celeb2 (1996 speakers)
  - CN-Celeb1 dev (797 speakers)
  - Combine the short utterances from the same speaker and genre to make each utterance longer than 5s

Duration(s)	CN-Celeb1		CN-Celeb2	
	# of Utters	Proportion	# of Utters	Proportion
<2	41,658	32.02%	36,505	6.89%
2-5	38,629	29.69%	57,215	10.81%
5-10	23,497	18.06%	266,799	50.39%
10-15	10,687	8.21%	154,120	29.11%
>15	15,638	12.02%	14,846	2.80%

Dataset	Speaker Num	Utterance Num
CN-Celeb2	1996	524787
CN-Celeb1 dev	797	126532
CN-Celeb2 Comb	1996	455603
CN-Celeb1 dev Comb	797	52693

Data Usage

- Training Dataset
  - CN-Celeb2 (1996 speakers)
  - CN-Celeb1 dev (797 speakers)
  - Combine the short utterances from the same speaker and genre to make each utterance longer than 5s
- Data Augmentation
  - Additive noise from MUSAN
  - Reverberation from RIR
  - Speed perturbation: speed up or slow down an utterance with ratio 1.1 and 0.9 (consider as a new speaker)

Data Usage

- Training Dataset
  - CN-Celeb2 (1996 speakers)
  - CN-Celeb1 dev (797 speakers)
  - Combine the short utterances from the same speaker and genre to make each utterance longer than 5s
- Data Augmentation
  - Additive noise from MUSAN
  - Reverberation from RIR
  - Speed perturbation: speed up or slow down an utterance with ratio 1.1 and 0.9 (consider as a new speaker)
- Feature Extraction
  - 80-dimensinal Fbank
  - No VAD is applied

System Architecture

Speaker Embedding Extractor

- TDNN based (x-vector, E-TDNN, ECAPA-TDNN)
- ResNet based (r-vector, Thin-resnet)
- Transformer based (s-vector)
- Embedding extractor from raw wav (RawNet, SincNet)

Complex and large models always have serious overfitting problem!

The simple r-vector doesn't seem to have this problem.

Snyder, David, et al. "X-vectors: Robust dnn embeddings for speaker recognition."
Snyder, David, et al. "Speaker recognition for multi-speaker conversations using x-vectors."
Desplanques, Brecht, et al. "Ecapa-tdnn: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification."
Hossein Zeinali, et al, “But system description to voxceleb speaker recognition challenge 2019”
Nagrani, Arsha, et al. "Voxceleb: Large-scale speaker verification in the wild." .
Katta, Sandesh V., et al. "S-vectors: Speaker embeddings based on transformer's encoder for text-independent speaker verification."
Jung, Jee-weon, et al. "Rawnet: Advanced end-to-end deep neural network using raw waveforms for text-independent speaker verification."
Ravanelli, Mirco, et al. "Speaker recognition from raw waveform with sincnet."

System Architecture

r-vector

Layer name	Structure	Output
Input	–	80 × Frame Num × 1
Conv2D-1	3 × 3, Stride 1	80 × Frame Num × 32
ResNetBlock-1	$\begin{bmatrix} 3\times3, 32 \\ 3\times3, 32 \end{bmatrix} \times 3$, Stride 1	80 × Frame Num × 32
ResNetBlock-2	$\begin{bmatrix} 3\times3, 64 \\ 3\times3, 64 \end{bmatrix} \times 4$, Stride 2	40 × Frame Num//2 × 64
ResNetBlock-3	$\begin{bmatrix} 3\times3, 128 \\ 3\times3, 128 \end{bmatrix} \times 6$, Stride 2	20 × Frame Num//4 × 128
ResNetBlock-4	$\begin{bmatrix} 3\times3, 256 \\ 3\times3, 256 \end{bmatrix} \times 3$, Stride 2	10 × Frame Num//8 × 256
StatsPooling	–	20 × 256
Flatten	–	5120
Emb Layer	–	256

Resnet34

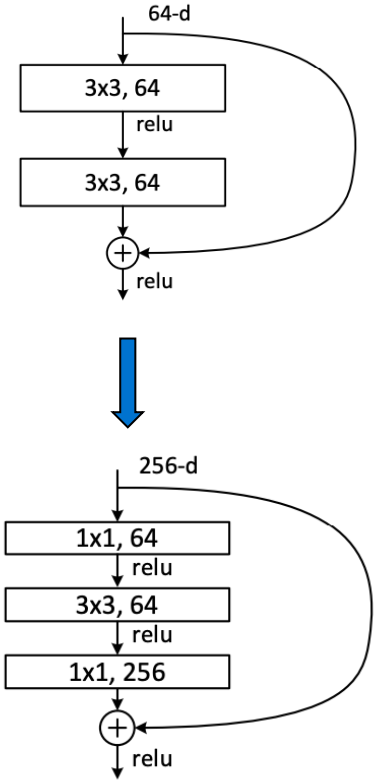
34-layer
7×7, 64, stride 2
3×3 max pool, stride 2
$\begin{bmatrix} 3\times3, 64 \\ 3\times3, 64 \end{bmatrix} \times 3$
$\begin{bmatrix} 3\times3, 128 \\ 3\times3, 128 \end{bmatrix} \times 4$
$\begin{bmatrix} 3\times3, 256 \\ 3\times3, 256 \end{bmatrix} \times 6$
$\begin{bmatrix} 3\times3, 512 \\ 3\times3, 512 \end{bmatrix} \times 3$

Hossein Zeinali, Shuai Wang, Anna Silnova, Pavel Matejka, and Old ˇ ˇrich Plchot, "But system description to voxceleb speaker recognition challenge 2019," arXiv preprint arXiv:1910.12592, 2019.

Make r-vector deeper

Layer name	Structure	Output
Input	–	80 × Frame Num × 1
Conv2D-1	3 × 3, Stride 1	80 × Frame Num × 32
ResNetBlock-1	$\begin{bmatrix} 1\times1, 32 \\ 3\times3, 32 \\ 1\times1, 128 \end{bmatrix} \times N_1$, Stride 1	80 × Frame Num × 128
ResNetBlock-2	$\begin{bmatrix} 1\times1, 64 \\ 3\times3, 64 \\ 1\times1, 256 \end{bmatrix} \times N_2$, Stride 2	40 × Frame Num//2 × 256
ResNetBlock-3	$\begin{bmatrix} 1\times1, 128 \\ 3\times3, 128 \\ 1\times1, 512 \end{bmatrix} \times N_3$, Stride 2	20 × Frame Num//4 × 512
ResNetBlock-4	$\begin{bmatrix} 1\times1, 256 \\ 3\times3, 256 \\ 1\times1, 1024 \end{bmatrix} \times N_4$, Stride 2	10 × Frame Num//8 × 1024
StatisticPooling	–	20 × 1024
Flatten	–	20480
Emb Layer	–	256

Deep ResNet Name	(N_1, N_2, N_3, N_4)
ResNet152	(3, 8, 36, 3)
ResNet221	(6, 16, 48, 3)
ResNet293	(10, 20, 64, 3)



System Training

- Stage I
  - Loss: Additive angular margin (AAM) loss
    - Margin: 0.2
    - Classification number: 8379 (2793 * 3)
  - Training Segment: 2s
  - Optimizer: SGD

$$L_3 = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{e^{s(\cos(\theta_{y_i}+m))}}{e^{s(\cos(\theta_{y_i}+m))}+\sum_{j=1,j\neq y_i}^{n}e^{s\cos\theta_j}}$$

- Stage II: Large Margin Finetune
  - Loss: Additive angular margin (AAM) loss
    - Margin: 0.5
    - Classification number: 2793 (no speed perturb aug)
  - Training Segment: 6s
  - Optimizer: SGD

System Scoring

- Scoring Method
  - Cosine + Asnorm (imposter cohort size 600)
- Methods to leverage multiple enrollment utterances
  - Utt-Concat: concatenate the utterances belonging to the same speaker to one long utterance
  - Emb-Avg: Average the embeddings of the enrollment utterances from the same speaker to one embedding
  - Score-Avg: Average the score between one test utterance and multiple enrollment utterances from the same speaker

Experiments

Ablation study on different ways to leverage multiple enrollment utterances

Scoring Method	Enroll Comb	minDCF (0.01)	EER (%)
Cosine	Utt-Concat	0.4391	7.305
Cosine	Emb-Avg	0.4004	6.922
Cosine + ASnorm	Utt-Concat	0.4035	7.085
Cosine + ASnorm	Emb-Avg	**0.3707**	**6.590**
Cosine + ASnorm	Score-Avg	0.4419	6.759

Note: results based one resnet34 after stage I training

- Asnorm benefit the system a lot
- Emb-Avg strategy achieves the best result

Experiments

The effect of speed perturbation

Model	Aug	minDCF (0.01)	EER
Resnet34	No Perturb	0.3958	7.981
Resnet34	With Perturb	0.3707	6.590

Note: results based on resnet34 after stage I training

- Speed perturbation is very necessary to improve the system performance

Experiments

Results for all the systems

System	Params #	minDCF (0.01)	EER (%)	FNR (%)	FPR (%)
ResNet34 *	6.63M	0.3958	7.981	35.29	0.043
ResNet34	6.63M	0.3707	6.590	31.73	0.054
ResNet152	19.8M	0.3386	5.762	29.34	0.045
ResNet221	23.8M	0.3270	**5.543**	28.08	0.046
ResNet293	28.6M	**0.3202**	5.553	**27.92**	**0.041**
DF-ResNet	14.8M	0.3361	6.279	28.83	0.048
ResNet34 + LM	6.63M	0.3543	6.221	30.06	0.054
ResNet152 + LM	19.8M	0.3251	5.452	28.66	0.039
ResNet221 + LM	23.8M	0.3179	5.284	28.27	**0.035**
ResNet293 + LM	28.6M	**0.3164**	**5.227**	27.82	0.038
DF-ResNet + LM	14.8M	0.3185	6.117	**27.46**	0.044
Fusion	-	0.2975	4.911	25.28	0.045

ResNet*: training without speed perturbation augmentation

FNR: false negative rate when we achieved the minDCF

FPR: false positive rate when we achieved the minDCF

LM: large margin finetuning

Bei Liu, Zhengyang Chen, Shuai Wang, Haoyu Wang, Bing Han, Yanmin Qian. DF-ResNet: Boosting Speaker Verification Performance with Depth-First Design10.21437/Interspeech.2020-2650. Accepted by InterSpeech 2022

Some other attempts

- TDNN based network

Model	Param#	minDCF (0.01)	EER
ECAPA-TDNN (C=512)	6.2M	0.4248	8.44
ECAPA-TDNN (C=1024)	14.7M	0.4343	8.345
ResNet34	6.63M	0.3543	6.221

Note: results after stage II training

- The ECAPA-TDNN doesn't perform well in this setup
- The ECAPA-TDNN seems to be overfitting

Desplanques, B., Thienpondt, J., Demuynck, K. (2020) ECAPA-TDNN: Emphasized Channel Attention, Propagation and Aggregation in TDNN Based Speaker Verification. Proc. Interspeech 2020, 3830-3834, DOI: 10.21437/Interspeech.2020-2650.

Some other attempts

- Leverage the genre information?
  - Adversarial training to remove the genre information
    - Add a gradient reversal layer before genre classification
    - No further improvement
    - Failed
  - Add the genre information in score calibration
    - Train acc: 0.9741 Valid acc: 71.75 (ECAPA-TDNN)
    - The EER and minDCF metric are not consistent (Especially under the CNSRC setup)
      - Score calibration objective is more related with EER
      - EER improved, minDCF degraded
    - Failed

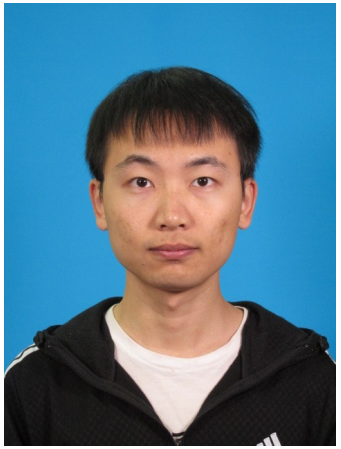
$$l(s) = w_s s + \mathbf{w}_q^T \mathbf{q} + b$$

Retrieval Task

- Directly submit the score of ResNet221

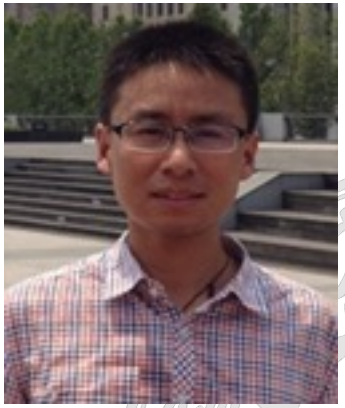
Zhengyang Chen

Bei Liu



Bing Han

Leying Zhang



Yanmin Qian

Thanks to our teams !