INTERSPEECH 2020
OCTOBER 25-29/ SHANGHAI , CHINA
SHANGHAI INTERNATIONAL CONVENTION CENTER

The xx205 System for the VoxCeleb Speaker Recognition Challenge 2020

Xu Xiang
AISpeech Ltd, China

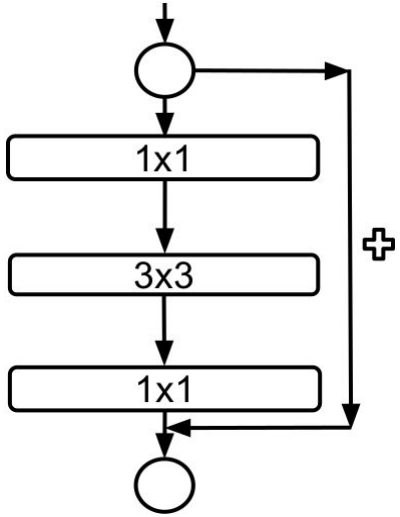
AISPEECH

Outline

- CNN Architectures for speaker modeling
- Composite margin loss for deep speaker verification
- Improved training strategies
- Score normalization and fusion
- System performance on VoxCeleb1, VoxSRC-19 and VoxSRC-20

ResNet architecture for deep speaker verification

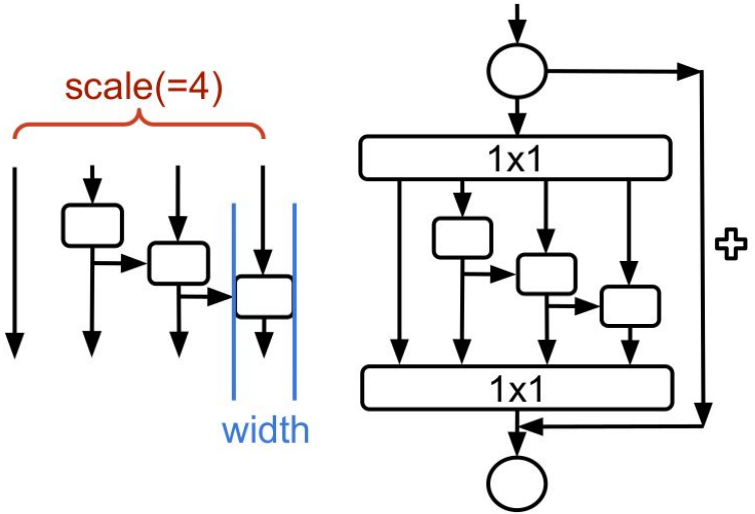
Layer	Kernel size	Stride	Output shape
Conv 1	$3 \times 3 \times 64$	1×1	$L \times 40 \times 64$
Blocks 1	$3 \times 3 \times 64$	1×1	$L \times 40 \times 64$
Blocks 2	$3 \times 3 \times 128$	2×2	$L/2 \times 20 \times 128$
Blocks 3	$3 \times 3 \times 256$	2×2	$L/4 \times 10 \times 256$
Blocks 4	$3 \times 3 \times 512$	2×2	$L/8 \times 5 \times 512$
StatPool	-	-	5×1024
Flatten	-	-	5120
Linear	-	-	256



ResNet Bottleneck block

Res2Net

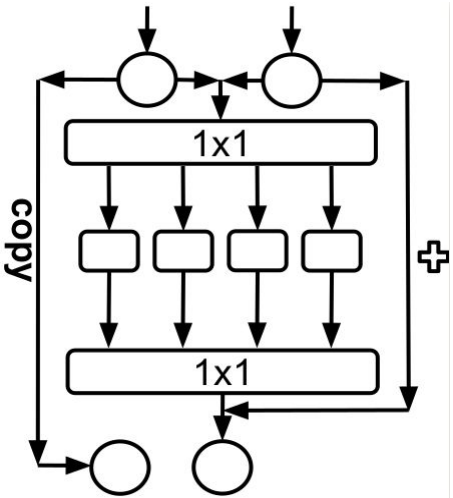
- 3x3 convolution in the ResNet bottleneck block is replaced by a series of smaller 3x3 convolutions
- Output features of the previous 3x3 convolution are sent to the next group of filters along with another group of input feature maps
- Easily plugged into existing CNN architectures



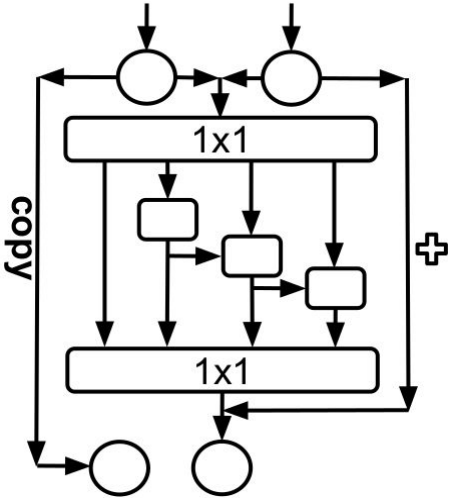
Res2Net block

Dual Path Networks (DPN)

- A DPN block contains a residual alike path (addition of features from different levels) and a densely connected alike path (concatenation of features from different levels)
- The Res2 structure can be incorporated into the DPN block



DPN block



Res2DPN block

Margin based losses for deep speaker verification

- Margin based loss encourages inter-class separability and intra-class compactness
- More discriminative on the decision boundary

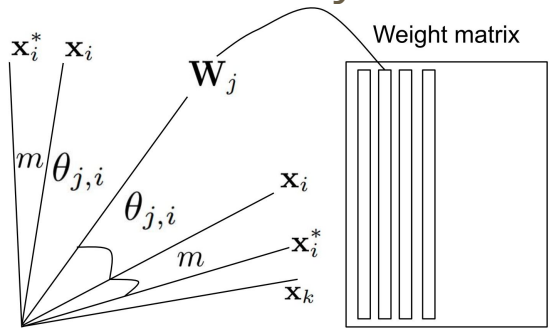
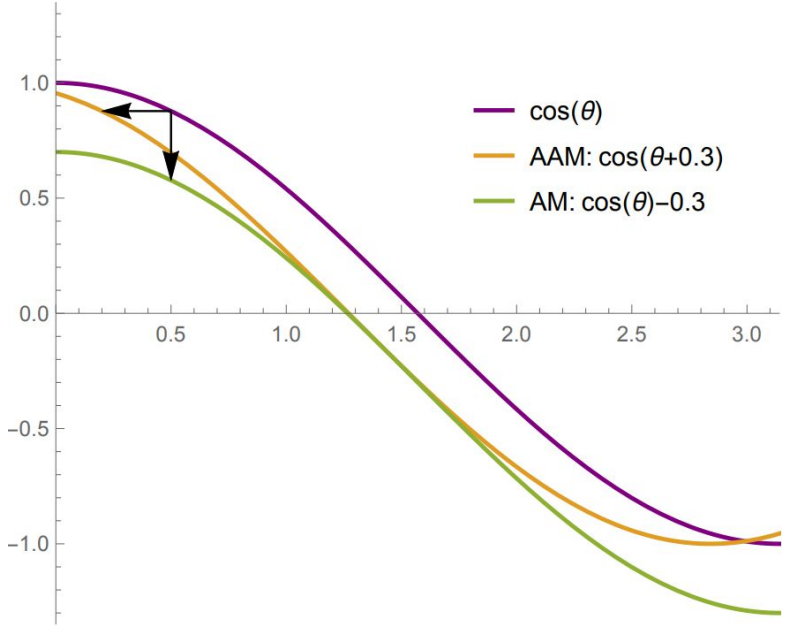


Illustration of the additive angular margin



Margin based losses: AAM and AM

Composite margin loss

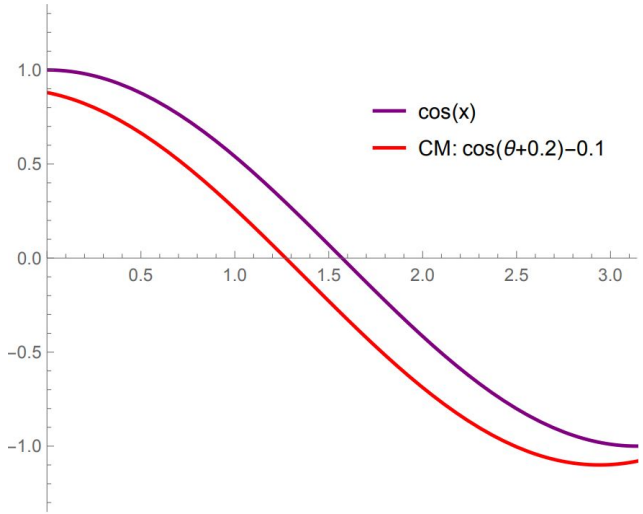
- The margin can be imposed by designing any function $\phi(\theta_{y_i,i})$ that satisfies

$$\phi(\theta_{y_i,i}) \leq \cos(\theta_{y_i,i})$$

- In this challenge, a composite margin loss is proposed:

$$\phi(\theta_{y_i,i}) = \cos(\theta_{y_i,i} + m_1) - m_2$$

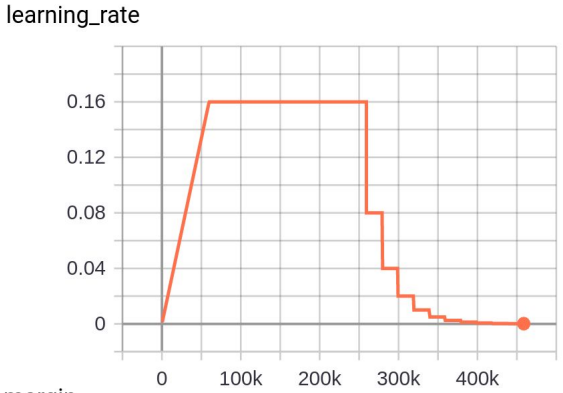
$$L_{\mathrm{CM}} = -\frac{1}{N}\sum_{i=1}^{N} \log \frac{e^{s(\cos(\theta_{y_i,i}+m_1)-m_2)}}{e^{s(\cos(\theta_{y_i,i}+m_1)-m_2)} + \sum_{j\neq i} e^{s\cos(\theta_{j,i})}}$$

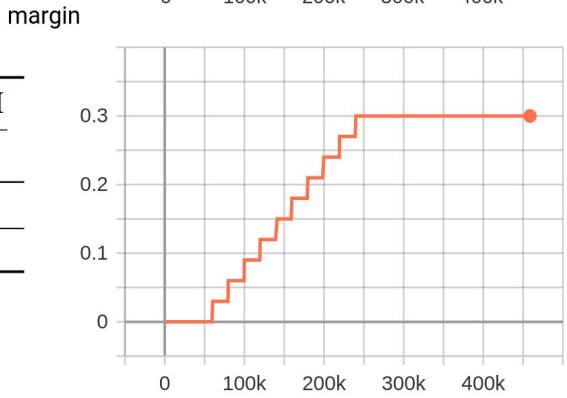


Margin based losses: CM

Neural network training

- 8 RTX 2080Ti GPUs, Tensorflow w/ Horovod
- LR/Margin warmup
- Train “sufficient” steps before decaying the LR
- Performance comparison between two training strategies:





Training strategy	Training data	Model	VoxCeleb1 EER(%)	VoxCeleb1-E EER(%)	VoxCeleb1-H EER(%)
previous	VoxCeleb2-dev (w/o Aug)	TDNN	2.694	2.762	4.732
current	VoxCeleb2-dev (w/o Aug)	TDNN	1.659	1.826	3.179

“Previous”: https://arxiv.org/abs/1906.07317

Score normalization and fusion

- Two sets of PLDA scores (based on VoxCeleb2-dev and VoxCelebCat) and a set of cosine scores are prepared for score normalization
- Adaptive symmetric normalization (adaptive s-norm) is adopted in this challenge (Cohort set: 5994 speakers in the VoxCeleb2 development set, top 400 cohorts are used)

$$s(e,t)_{as\text{-}norm1} = \frac{1}{2} \cdot \left(\frac{s(e,t) - \mu(S_e(\mathcal{E}_e^{top}))}{\sigma(S_e(\mathcal{E}_e^{top}))} + \frac{s(e,t) - \mu(S_t(\mathcal{E}_t^{top}))}{\sigma(S_t(\mathcal{E}_t^{top}))} \right)$$

- Bosaris toolkit is used to train a linear fuser to do the score fusion

Results on VoxCeleb1 test sets and VoxSRC-20 validation set

- Cosine scores after adaptive s-norm are used for performance evaluation
- Among all three architectures, DPN performs the best

Track	Model	VoxCeleb1		VoxCeleb1-E		VoxCeleb1-H		VoxSRC-20 Val	
		EER(%)	MinDCF	EER(%)	MinDCF	EER(%)	MinDCF	EER(%)	$\text{MinDCF}_{0.05}$
	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
1	Res2Net50	0.8137	0.0940	0.8968	0.0967	1.693	0.1536	2.896	0.1473
	SE-ResNet34	0.8457	0.0780	0.9875	0.1000	1.748	0.1631	3.029	0.1508
	ResNeXt50	0.8882	0.1104	1.025	0.1095	1.855	0.1678	3.125	0.1561
	DPN101	0.7925	0.0782	0.9113	0.0961	1.67	0.1558	2.822	0.1471
	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
2	DPN50	0.8457	0.0799	0.9523	0.1008	1.669	0.1594	2.873	0.1467
	DPN68	0.8137	0.0801	0.8806	0.0959	1.591	0.1512	2.731	0.1427

Results on VoxSRC-20 evaluation set

- In track 1, the submitted systems consists 24 subsystems of DPN, Res2Net and ResNet trained on VoxCeleb2 development set
- In track 2, two additional subsystems of DPN trained on VoxCeleb2 development set and Librispeech data
- The track 1 fusion system can achieve very strong result on VoxSRC-19 evaluation set

Track	VoxSRC-19 Eval	VoxSRC-20 Val		VoxSRC-20 Eval	
	EER(%)	EER(%)	$\text{minDCF}_{0.05}$	EER(%)	$\text{minDCF}_{0.05}$
1	0.6868	1.866	0.0982	3.808	0.1958
2	-	1.818	0.0979	3.798	0.1942

Thank you!