Advances in Applied Mathematics 应用数学进展, 2021, 10(11), 3883-3889
Published Online November 2021 in Hans. http://www.hanspub.org/journal/aam
https://doi.org/10.12677/aam.2021.1011412

Hans汉斯

基于余弦Softmax损失的说话人验证研究

谢旗旺，周林华*

长春理工大学，数学与统计学院，吉林 长春

收稿日期：2021年10月17日；录用日期：2021年11月7日；发布日期：2021年11月22日

摘　要

基于神经网络提取说话人嵌入在说话人验证任务上显示出了良好的性能，然而传统的说话人嵌入网络通常采用Softmax损失作为训练标准，其说话人特征类间区分不明显。因此本文通过引入负样本对(来自不同类别的两个样本)学习改进Softmax损失，在余弦角度空间中学习不同说话人特征存在明显角度间距且同一说话人特征聚集紧密的嵌入特征。在公开数据集AISHELL数据集上进行的说话人验证实验表明，与A-softmax相比，该损失函数的等误差率更低且ROC曲线下面积更大。

关键词

Softmax损失，说话人嵌入，说话人验证，角间距

Speaker Verification Analysis Based on Cosine Softmax Loss

Qiwang Xie, Linhua Zhou*

School of Mathematics and Statistics, Changchun University of Science and Technology, Changchun Jilin

Received: Oct. 17th, 2021; accepted: Nov. 7th, 2021; published: Nov. 22nd, 2021

Abstract

Neural network-based extraction of the speaker embedding in the speaker verification task shows good performance, however, the traditional speaker embedding network usually uses Softmax loss as the loss function, and the distinction between the speaker characteristics class is not obvious. Therefore, this paper improves Softmax loss by introducing negative sample pairs (two samples from different categories) to learn speaker embedding in cosine angle space that there is a clear angle distance between different speakers and that the same speaker features are compact.

*通讯作者。

文章引用: 谢旗旺, 周林华. 基于余弦 Softmax 损失的说话人验证研究[J]. 应用数学进展, 2021, 10(11): 3883-3889.
DOI: 10.12677/aam.2021.1011412

The speaker verification experiment on the public data set AISHELL dataset showed that the loss function had a lower EER and a larger area under the ROC curve than A-softmax.

Keywords

Softmax Loss, Speaker Embedding, Speaker Verification, Angle Distance

1. 引言

说话人验证是从语音中提取出说话人特征信息，以确定两个语音段是否来自同一说话人。深度学习的兴起之前，提取说话人信息特征的方法主要是高斯混合模型(GMM)、i-vector 等算法[1] [2] [3] [4]。随着深度学习的兴起，由于其具有强大的自主学习能力以及充分挖掘特征潜在信息的能力，在说话人识别领域取得了优异的性能表现，并成为了主流的声纹特征提取方法[5] [6] [7] [8]。深度神经网络在用于声纹特征提取时，损失函数作为待优化的目标函数，极大影响了所学声纹特征的判别性，因此，许多学者在损失函数的设计上展开了深入研究。

其中 Softmax 损失是分类网络中最常见的损失函数，也是训练说话人嵌入神经网络最常用的损失函数之一，但它学习的特征区分性不够明显，具体表现为不同类特征间的距离不够远。因此，人们提出了几种 Softmax 的改进损失来增强特征的类间区分性，包括 L-softmax [9]、A-softmax [10]、SpherFace [11]、CosFace [12]和 ArcFace [13]等等，它们也被引入到说话人识别领域中[14] [15] [16]。这些方法主要通过将特征和分类权重向量单位化投影到超球面上，使分类结果取决于特征和分类权重向量之间的角度，再对类间间隔施加一个惩罚因子，使不同说话人特征的间距更加明显。然而，超参数间距惩罚因子的设置又是一个新的难点。基于深度度量学习的端到端损失如对比损失[8]和三元组损失[17]直接优化说话人嵌入间的距离。其中对比损失以成对样本进行训练，使得同一类样本对的特征距离缩小，不同类样本对特征之间的距离拉开，因此对比损失能学习到具有明显类间距离的特征表示。受此启发，本文在分类神经网络的训练过程中引入非同类样本对的学习，通过对非同类样本对特征距离的优化拉大样本的类间距离。

此外，与欧几里得距离相比，角距离是特征空间中更自然的选择，因此，本文沿袭 Softmax 改进损失的思想将特征和分类权重向量单位化到超球面上，再通过优化不同类样本对之间的余弦值来增加特征的类间角度距离，使分类权重向量彼此远离，该损失函数称为余弦 Softmax 损失。在 AISHELL 语音数据集上的说话人验证实验结果表明，余弦 Softmax 损失的表现要优于 Softmax 损失以及其改进损失 A-softmax 损失。

2. 具有角度判别性的深度说话人嵌入

2.1. 余弦 Softmax 损失

在多分类的情况下神经网络的 Softmax 损失为：

$$L_1 = -\frac{1}{k}\sum_{i}^{k}\log\frac{e^{w_{y_i}^{T}x_i+b_{y_i}}}{\sum_j e^{w_j^{T}x_i+b_j}}.$$

其中 k 是样本数量，w 是最后层分类层的权重项，b 是分类层的偏置项，x_i 是第 i 个训练样本，w_j, w_{y_i} 分别是 w 的第 j 列和第 y_i 列，b_j, b_{y_i} 分别是 b 的第 j 项和第 y_i 项。因为 $w_j \cdot x_i = \|w_j\|\|x_i\|\cos\theta$，$\theta$ 是 w_j 与 x_i 之间的夹角，所以 Softmax 损失可以写为：

$$L_1 = -\frac{1}{k}\sum_i^k \log \frac{e^{\|w_{y_i}^{\mathrm{T}}\|\|x_i\|\cos(\theta_{y_i,i})+b_{y_i}}}{\sum_j e^{\|w_j^{\mathrm{T}}\|\|x_i\|\cos(\theta_{j,i})+b_j}}$$

$\theta_{y_i,i}$ 是 w_{y_i} 与 x_i 之间的夹角。

对 Softmax 施加额外约束，令权重范数 $\|w\|=1$ 且 $\|x_i\|=1$，并去掉偏置项 b，则损失为：

$$L_2 = -\frac{1}{k}\sum_i^k \log \frac{e^{\cos(\theta_{y_i,i})}}{\sum e^{\cos(\theta_{j,i})}}.$$

设 x_i', x_j' 是从网络最后隐藏层特取样本 x_i, x_j 的特征向量，且 x_i, x_j 不属于同类样本，则该对样本特征的相似度 D 为：

$$D = x_i' \cdot x_j' = \|x_i'\| \cdot \|x_j'\| \cos\left(\theta_{x_i', x_j'}\right),$$

$$令 \|x_i'\| = 1, \|x_j'\| = 1,$$

$$\square D = \cos\left(\theta_{x_i', x_j'}\right).$$

给定两个负样本对的相似度阈值 α，$1 > \alpha > 0$，优化每对特征相似度小于阈值的负样本，使得负样本对特征都能有不小于 $\arccos(-\alpha)$ 的角度距离，取 $\alpha = 0$，则两个不同类样本至少存在垂直的角度间距，因此定义负样本对学习的损失函数为：

$$L_3 = \left[\max\left(0, D+\alpha\right)\right]^2.$$

余弦 Softmax 是 L_2 损失和 L_3 损失的和，λ 用于平衡负样本对的损失：

$$L_4 = L_2 + \lambda L_3.$$

其解释如图 1 所示，L_2 损失以分类权重向量为中心聚拢同类样本特征，L_3 损失通过负样本对之间的距离优化，带动分类权重向量彼此分隔开，这种联合效应可以学习到更有判别性的说话人嵌入。

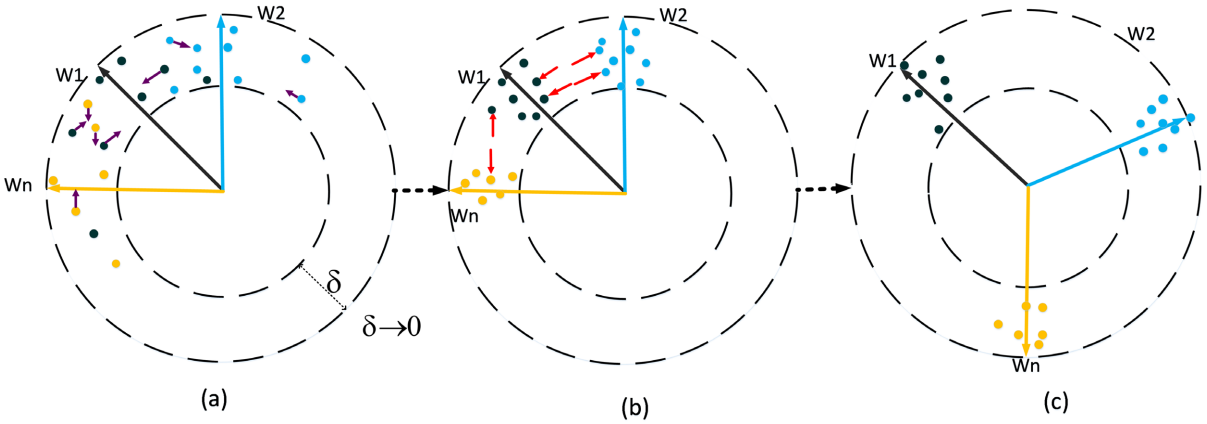

Figure 1. Geometric interpretation of cosine Softmax loss
图 1. 余弦 Softmax 损失的几何解释

2.2. 网络模型设计

通过设计两个共享权重参数的特征提取网络组成孪生网络进行训练。特征提取网络又由卷积层Conv1、残差块[18]、均值池化层Avg-pool、全连接层和分类层组成。网络训练完成后从最后的隐藏层提取语音的声纹特征，即：

$$x_i' = Net(x_i), x_j' = Net(x_j).$$

x_i, x_j 为说话人语音，经过特征提取网络 Net 后得到深度说话人特征 x_i' 和 x_j'。

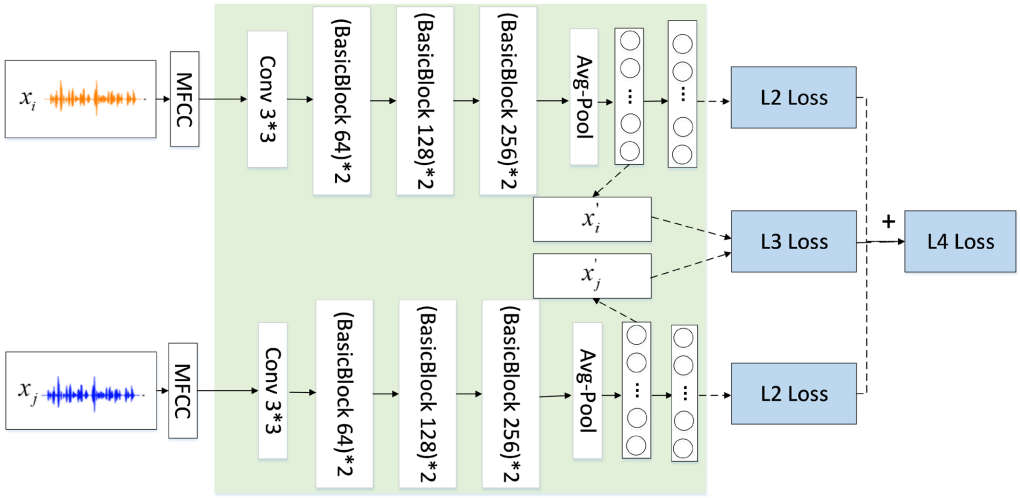

Figure 2. Training structure of network
图 2. 网络训练结构图

网络训练结构图如图 2 所示，通过最小化以下损失函数来优化网络参数：

$$\arg\min_{Net(\cdot)} L_3 = -\frac{1}{2n}\left(\sum_{i=1}^{n}\log\frac{e^{\cos\left(\theta_{x_i',w_{y_i}}\right)}}{\sum_k^K e^{\cos\left(\theta_{x_i',w_k}\right)}} + \sum_{j=1}^{n}\log\frac{e^{\cos\left(\theta_{x_j',w_{y_j}}\right)}}{\sum_k^K e^{\cos\left(\theta_{x_j',w_k}\right)}}\right) + \frac{\lambda}{n}\sum_{l=1}^{n}\left[\max\left(0,\alpha - D\right)\right]^2$$

$$y_i \neq y_j, \alpha = 0, \lambda = 1.$$

$y_i \neq y_j$ 表示两条语音不属于同一说话人，n 为训练的样本对数，K 表示分类的说话人数量。

3. 实验及结果分析

3.1. 数据集

实验采用 AISHE 数据集[19]，该数据集是公开的最大的中文语音数据集，从中选取 100 个说话人，每个说话人选取 100 条语音作为训练集语音，20 条语音作为测试集语音。

3.2. 实验设置

输入特征：用基于能量的语音活动检测方法去除沉默语音段，采用 32 ms 长度汉明窗和 16 ms 的窗口移位提取语音的 MFCC 特征。

网络设置：训练说话人嵌入的孪生网络参数如表 1 所示。前端提取器基于 ResNet [18]架构，ReLU 激活函数和批归一化应用于每个卷积层，Dropout 层设置沉默的神经元比例为 0.4，全连接层设置 64 个神

经元，提取 64 维的说话人嵌入特征，最后的分类层 100 个节点对应训练数据的 100 个说话人，具体网络结构如下表。

Table 1. The structure parameters of speaker embedding network
表 1. 说话人嵌入网络结构参数

Layer	Structure	Output size
Input	—	$300\times39\times1$
Conv1	3×3,Stride 1	$300\times39\times64$
BasicBlock 1	$\begin{bmatrix}3\times3,64\\3\times3,64\end{bmatrix}\times2$	$300\times39\times64$
BasicBlock 2	$\begin{bmatrix}3\times3,128\\3\times3,128\end{bmatrix}\times2$	$300\times39\times128$
BasicBlock 3	$\begin{bmatrix}3\times3,256\\3\times3,256\end{bmatrix}\times2$	$300\times39\times256$
Avg_pool	—	256×4
Dropout	—	—
Dense	1024×64	64
Dense	64×100	100

3.3. 实验结果

从测试集根据正负样本对 1:1 的比例抽取 10,000 对测试样本，采用 ROC 曲线下面积(AUC)和等误差率(EER)来评估结果，不同损失函数训练的实验结果如表 2 所示，ROC 曲线图如图 3 所示。

Table 2. Speaker verification results of different loss functions
表 2. 不同损失函数的说话人验证实验结果

损失函数	AUC	EER/%
Softmax	0.970	7.04
A-Softmax	0.976	5.74
CosineSoftmax	**0.983**	**4.88**

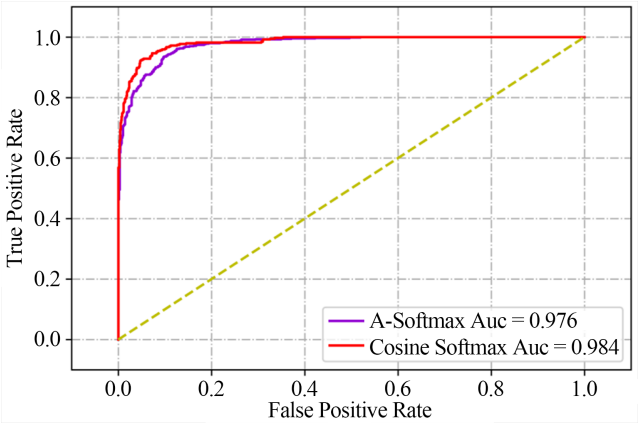

Figure 3. ROC curves of cosine Softmax and A-softmax
图 3. 余弦 Softmax 和 A-softmax 损失的 ROC 曲线图

由实验结果可知，在相同网络结构和数据集的条件下，改进损失函数余弦 Softmax 损失函数的实验结果优于 Softmax 和 A-softmax，取得了最低的 EER 4.88%，同时该损失测试的 AUC 最大为 0.983。

4. 总结

本文的余弦 Softmax 损失以引入负样本对学习的方式对 Softmax 损失进行了改进，有效改善了 Softmax 损失类间距离不显著的问题，并在说话人验证实验中取得了优于 A-softmax 损失的性能。

基金项目

国家自然科学基金(11426045)、吉林省自然科学基金学科布局项目(20180101229JC)。

参考文献

[1] Reynolds, D.A., Quatieri, T.F. and Unn, D.R. (2000) Speaker Verification Using Adapted Gaussian Mixture Models. *Digital Signal Processing*, **10**, 19-41. https://doi.org/10.1006/dspr.1999.0361

[2] Trabelsi, I., Ayed, D.B. and Ellouze, N. (2016) Comparison between GMM-SVM Sequence Kernel and GMM: Application to Speech Emotion Recognition. *Journal of Engineering Science and Technology*, **11**, 1221-1233.

[3] Kanagasundaram, A., Vogt, R., Dean, D., *et al*. (2011) i-Vector Based Speaker Recognition on Short Utterances. *INTERSPEECH*, Florence, 27-31 August 2011, 2341-2344. https://doi.org/10.21437/Interspeech.2011-58

[4] Dehak, N., Kenny, P.J., Dehak, R., *et al*. (2011) Front-End Factor Analysis for Speaker Verification. *IEEE Transactions on Audio Speech and Language Processing*, **19**, 788-798. https://doi.org/10.1109/TASL.2010.2064307

[5] Snyder, D., Garcia-Romero, D., Povey, D., *et al*. (2017) Deep Neural Network Embeddings for Text-Independent Speaker Verification. *INTERSPEECH*, Stockholm, 20-24 August 2017, 999-1003. https://doi.org/10.21437/Interspeech.2017-620

[6] Snyder, D., Garcia-Romero, D., Sell, G., *et al*. (2018) X-Vectors: Robust DNN Embeddings for Speaker Recognition. Proc. *ICASSP*, Calgary, 15-20 April 2018, 5329-5333. https://doi.org/10.1109/ICASSP.2018.8461375

[7] Variani, E., Lei, X., Mcdermott, E., *et al*. (2014) Deep Neural Networks for Small Footprint Text-Dependent Speaker Verification. *IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*), Florence, 4-9 May 2014, 4052-4056. https://doi.org/10.1109/ICASSP.2014.6854363

[8] Heigold, G., Moreno, I., Bengio, S., *et al*. (2016) End-To end Text-Dependent Speaker Verification. *IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*), Shanghai, 20-25 March 2016, 5115-5119. https://doi.org/10.1109/ICASSP.2016.7472652

[9] Liu, W., Wen, Y., Yu, Z., *et al*. (2016) Large-Margin Softmax Loss for Convolutional Neural Networks. *The* 33*rd International Conference on Machine Learning* (*ICML* 2016), New York, 19-24 June 2016, 507-516.

[10] Liu, W., Wen, Y., Yu, Z., *et al*. (2017) Sphereface: Deep Hypersphere Embedding for Face Recognition. *The IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), Honolulu, 21-26 July 2017, 212-220. https://doi.org/10.1109/CVPR.2017.713

[11] Wang, F., Xiang, X., Cheng, J., *et al*. (2017) NormFace: L2 Hypersphere Embedding for Face Verification. *Proceedings of the* 25*th ACM International Conference on Multimedia*, Mountain View, 23-27 October 2017, 1041-1049. https://doi.org/10.1145/3123266.3123359

[12] Wang, H., Wang, Y., Zhou, Z., *et al*. (2018) CosFace: Large Margin Cosine Loss for Deep Face Recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Salt Lake City, 18-23 June 2018, 5265-5274. https://doi.org/10.1109/CVPR.2018.00552

[13] Wang, F., Liu, W., Liu, H., *et al*. (2018) Additive Margin Softmax for Face Verification. *IEEE Signal Processing Letters*, **25**, 926-930. https://doi.org/10.1109/LSP.2018.2822810

[14] Huang, Z., Wang, S. and Yu, K. (2018) Angular Softmax for Short Duration Text-Independent Speaker Verification. *INTERSPEECH*, Salt Lake City, 18-23 June 2018, 3623-3627. https://doi.org/10.21437/Interspeech.2018-1545

[15] Yu, Y.Q., Fan, L. and Li, W.J. (2019) Ensemble Additive Margin Softmax for Speaker Verification. 2019 *IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*), Brighton, 12-17 May 2019, 6046-6050. https://doi.org/10.1109/ICASSP.2019.8683649

[16] Li, Y., Gao, F., Ou, Z., *et al*. (2019) Angular Softmax Loss for End-to-End Speaker Verification. 2018 11*th International Symposium on Chinese Spoken Language Processing* (*ISCSLP*), Taipei, 26-29 November 2018, 190-194. https://doi.org/10.1109/ISCSLP.2018.8706570

[17] Bredin, H. (2017) Tristounet: Triplet Loss for Speaker Turn Embedding. *IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*), New Orleans, 5-9 March 2017, 5430-5434. https://doi.org/10.1109/ICASSP.2017.7953194

[18] He, K., Zhang, X., Ren, S., *et al*. (2016) Deep Residual Learning for Image Recognition. *IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, 27-30 June 2016, 770-778. https://doi.org/10.1109/CVPR.2016.90

[19] Bu, H., Du, J., Na, X., *et al*. (2017) AISHELL-1: An Open-Source Mandarin Speech Corpus and a Speech Recognition Baseline. *O-COCOSDA*, Seoul, 1-3 November 2017, 1-5. https://doi.org/10.1109/ICSDA.2017.8384449