Text-Independent Speaker Verification Based on Triplet Convolutional Neural Network Embeddings

一. 摘要

The effectiveness of introducing deep neural networks into conventional speaker recognition pipelines has been broadly shown to benefit system performance. A novel text-independent speaker verification(SV) framework based on the triplet loss and a very deep convolutional neural network architecture (i.e., Inception-Resnet-v1) are investigated in this study, where a fixed length speaker discriminative embedding is learned from sparse speech features and utilized as a feature representation for the SV tasks. A concise description of the neural network-based speaker discriminative training with triplet loss is presented. An Euclidean distance similarity metric is applied in both network training and SV testing, which ensures the SV system to follow an end-to-end fashion. By replacing the final max/average pooling layer with a spatial pyramid pooling layer in the Inception-Resnet-v1 architecture, the fixed-length input constraint is relaxed and an obvious performance gain is achieved compared with the fixed-length input speaker embedding system. For datasets with more severe training/test condition mismatches, the probabilistic linear discriminant analysis (PLDA) back end is further introduced to replace the distance-based scoring for the proposed speaker embedding system. Thus, we reconstruct the SV task with a neural network based front-end speaker embedding system and a PLDA that provides channel and noise variabilities compensation in the back end. Extensive experiments are conducted to provide useful hints that lead to a better testing performance. Comparison with the state-of-the-art SV frameworks on three public datasets (i.e., a prompt speech corpus, a conversational speech Switchboard corpus, and NIST SRE10 10 s–10 s condition) justifies the effectiveness of our proposed speaker embedding system

将深度学习引入语音识别已经展现出不错的性能改进。本文研究了一种基于 triplet loss 和深层卷积神经网络的新型文本无关语音识别（SV）框架，这

个框架从稀疏语音特征中学习固定长度的 embeddings，作为语音识别任务的特征表示。我们给出了基于 triplet loss 的卷积神经网络训练的简要描述。将欧几里得相似性度量应用于网络训练和 SV 测试，以确保 SV 系统遵循端到端的方式。通过在 Inception-Resnet-v1 架构中，用金字塔池化代替传统的最大池化，可以减小固定输入长度的约束，并获得明显的性能提高。为了处理训练/测试条件不匹配的数据集，我们引入了概率线性判别分析（PLDA）来代替基于距离的打分系统。因此，我们使用基于深度学习的语音特征嵌入系统作为前段，拥有通道噪声可变性补偿的 PLDA 作为后端来重构语音识别系统。大量的实验提供了有用的参数，使得我们的系统获得了更好的性能。与三个公共数据集（提示语语音库、会话语音库、NIST）上最新的研究成果相比，我们的系统，我们语音特征嵌入框架拥有更好的表现

二． 精读论文

2.1 基本识别框架

声纹识别系统的基本框架如图一所示：整个系统分为训练阶段和识别阶段。在训练阶段，首先对训练语音进行预处理和特征提取，然后构建能代表说话人的

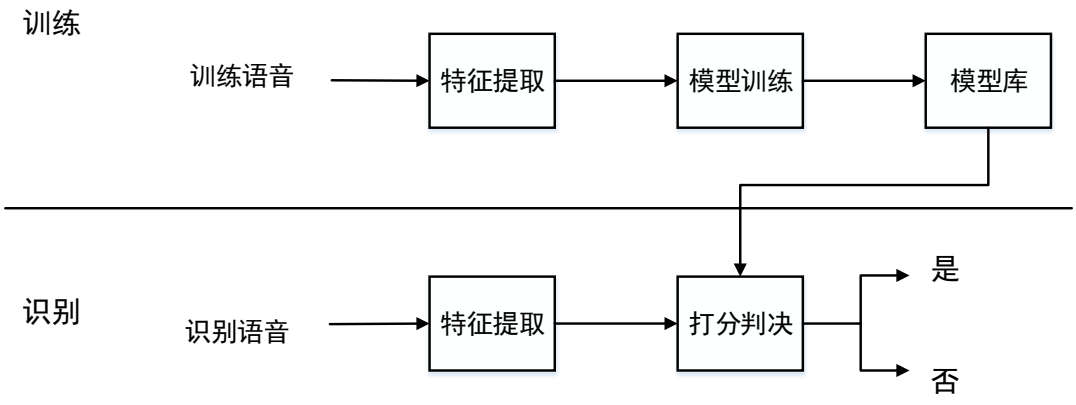

图 1 声纹识别基本框架

模型将其存放在模型库中。在识别阶段，同样需要对待识别的语音进行预处理、特征提取和模型构建，然后把得到的模型与模型库中的模型进行一对一的相似性比较，判断是否属于同一个人。

2.2 模型及 triplet loss

论文的模型为 triplet loss，如公式一所示。图 2 为 triplet 三元组的示意图。Triplet loss: 首先选择一个样本 Anchor，Anchor 属于身份 A，再选择一个样本 Positive，同样属于 A，然后再选择一个样本 Negative，该样本不属于 A，triplet loss 的目的就是使得属于同一个身份的样本之间的距离尽可能地近，不同身份样本之间的距离尽可能的远，最终得到能够代表一个说话人的 embedding 向量。

$$\Delta_i = \|f(x_i^a) - f(x_i^p)\|_2^2 - \|f(x_i^a) - f(x_i^n)\|_2^2 + \alpha,$$
$$L = \sum_{i=1}^{N} \max(0, \Delta_i), (x_i^a, x_i^p, x_i^n) \in \mathcal{T} \quad (1)$$

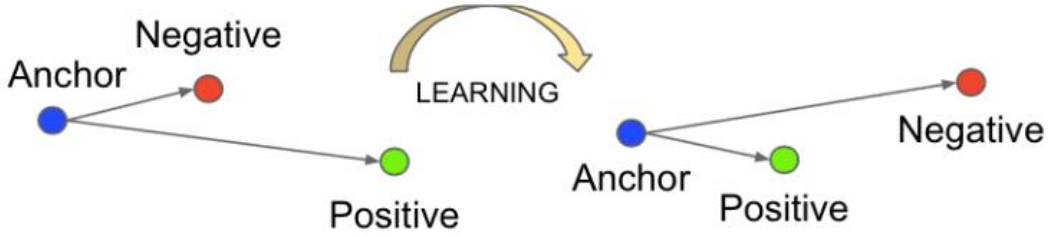

图 2 triplet 三元组示意图

2.3 论文系统方案

论文的系统方案如图 3 所示。系统分为 3 个阶段：

训练阶段

（1）. 从数据集中每次选取一个 batch 的 triplet 三元组（anchor、positive、negative）

（2）. 对每条语音提取 fbank 特征或 FFT 频谱特征，每条语音得到一个二维的特征矩阵。

（3）. 把一个三元组的每个特征矩阵分别输入到网络中，得到输出结果，输出为 128 维向量。

（4）. 对输出的结果进行 L2 归一化处理。

（5）. 计算一个三元组的 triplet loss，并以同样的方式计算出一个 batch 三元组的 triplet loss

（6）.根据 triplet loss 计算出梯度并更新网络模型的参数。

注册阶段：

（1）. 采集说话人的语音

（2）. 对语音进行有效性检测，去除静音并分成 4s 的定长语音片段

（3）. 分别提取语音片段的 fbank 特征或者 FFT 变换的频谱特征。

（4）. 把提取的特征矩阵输入到训练好的网络中得到多个 128 维的向量

（5）. 对输出的向量分别做 L2 归一化，然后取平均值，再进行 L2 归一化，得到代表一个说话人的 embedding，并存入数据库中。

测试阶段：

（1）.按照与注册阶段相同的过程，得到该语音的 embedding

（2）.把该条 embedding 分别与数据库中的 embedding 做相似性比较，判断是否属于同一个人

三. 论文复现

3.1 实现流程

实现流程如图 3 所示，分为三个部分：

1. **数据预处理**：首先对数据集中的样本进行语音有效性检测（VAD）去除静音部分，然后将语音样本 进行切割，长度为 4 秒，滑动的 step 为 2 秒。

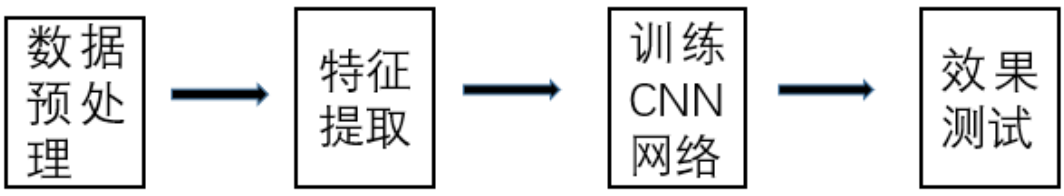

图 3 实现流程图

2.特征提取：输入特征采用两种方式：Fbank 特征或 FFT 特征。经过预处理后的语音采样率为 8K，然后进行分帧，窗口长度为 0.032 秒，step 为 0.016 秒，4 秒长的语音段会被分成 250 帧。对于 Fbank 特征，每帧提取 120 维的特征，4 秒的

语音形成 120*250 的特征矩阵；对于 fft 特征，对每帧数据做256个点的FFT,取前128个频谱分量作为特征,4秒的语音就构成128*250的特征矩阵。

3.训练CNN：构建 triplet 三元组，把 triplet loss 作为优化函数，对 CNN 网络进行训练。

3.2 triplet 三元组选取

三元组的种类可以分为下列三种，如图4所示：easy triplets：可以使loss=0的三元组；hard triplets：d(a,n) < d(a,p)的三元组，即一定会误识别的三元组；semi-hard triplets：d(a,p) < d(a,n) <d(a,p) + margin 的三元组，即处在模糊区域的三元组。

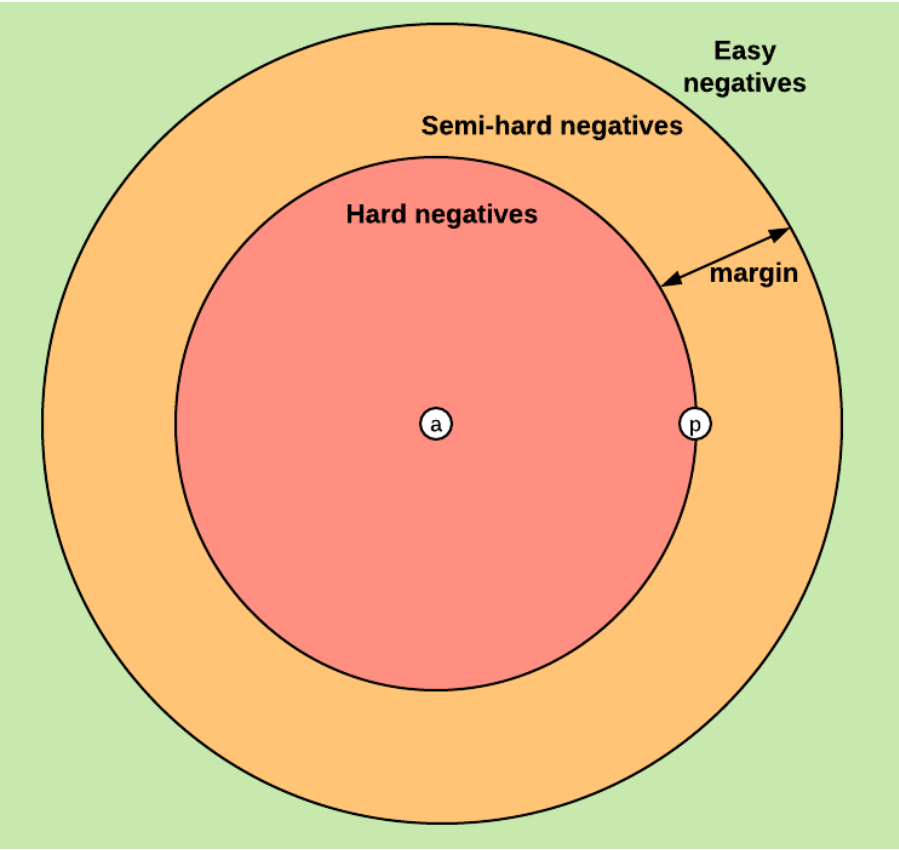

图4 三元组种类示意图

三元组的选取有离线方式和在线方式。离线方式：可以在每轮迭代之前从所有triplet 中选择 semi-hard Triplet。也就是先对所有的训练集计算嵌入表达(feature)，然后只选择semi-hard triplets 并以此为输入训练一次网络。因为

每轮训练迭代之前都要遍历所有 triplet，计算它们的嵌入，所以 offline 挖掘 triplet 效率很低；在线方式：假设有 B 个图片（不是 Triplet），也就是可以生成 B 个嵌入表达，那么我们最多以此生成 B 的三次方个 Triplet，当然大多数 Triplet 都不符合要求，优点是只用遍历 B 个图片，效率高。

3.3 实验及结果

数据集采用的是 NIST SRE16 数据集，训练使用 600 个人，每个人 1 到 3 条语音，每条语音 1 分钟左右。测试时用 200 个人进行注册，每个人挑出 5 条长度维 8 秒的语音段进行测试。实验共 200*5*200=200000 个。网络结构如图 5 所示，实验结果如图 6 所示，等错误率(EER)为 5.21%。

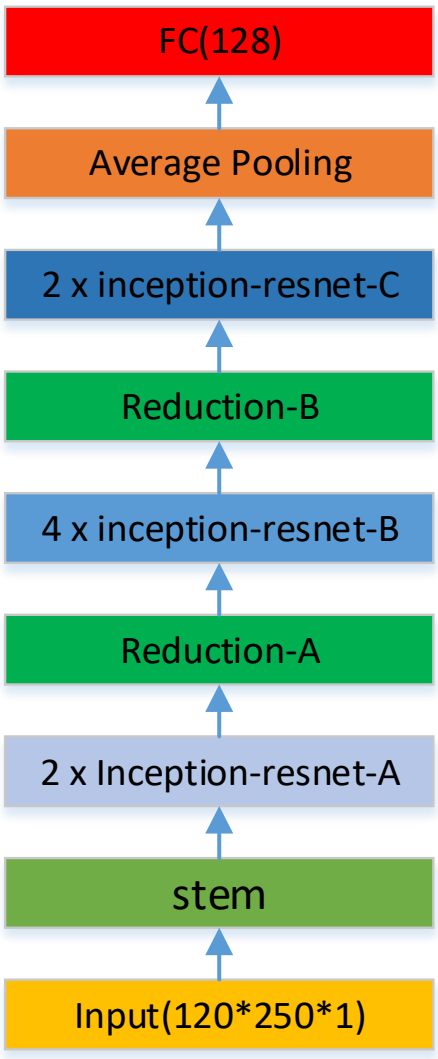

图 5 网络结构

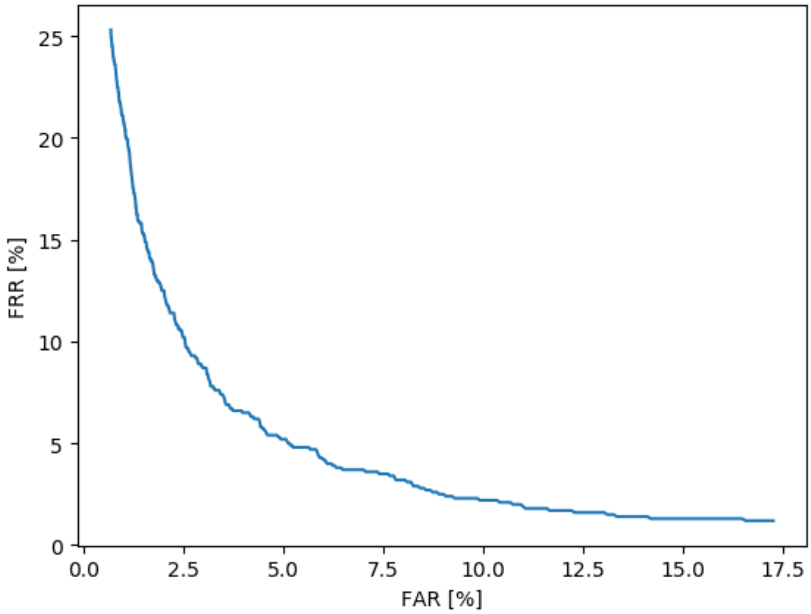

图 6 实验结果图

在网络训练的过程中损失函数的值会逐渐下降，但是不会收敛，因为每次生成的 triplet 三元组都不同，loss 的图像如图 7 所示。但是每个 batch 中不满足 loss 等于 0 的三元组的比例会逐渐下降。如图 8 所示。

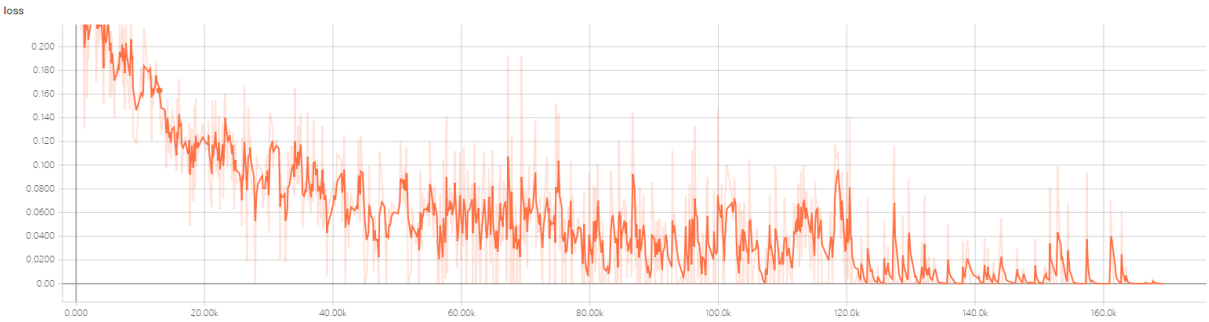

图 7 loss 图像

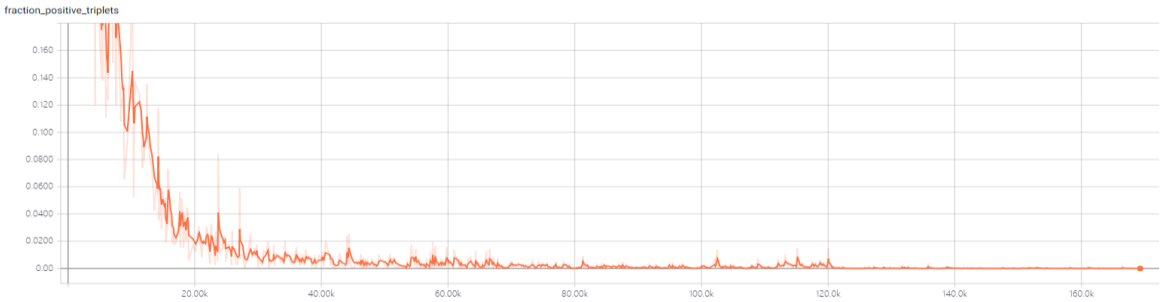

图 8 fraction_positive_triplets 的图像

四． 对比实验

论文中使用的是 SRE10 数据集，EER 最好可达 2.97%，但是 SRE10 数据集不可获取，并且论文中并没有给出实验的具体过程和设置，所以无法进行比较。

五． 结论

使用 triplet loss 和神经网络来实现声纹识别，具有中间过程比较少的优点，准确率与传统 i-vector 方法相近，甚至在长语音下更加优秀。但在较短的语音下，效果还有待提高。