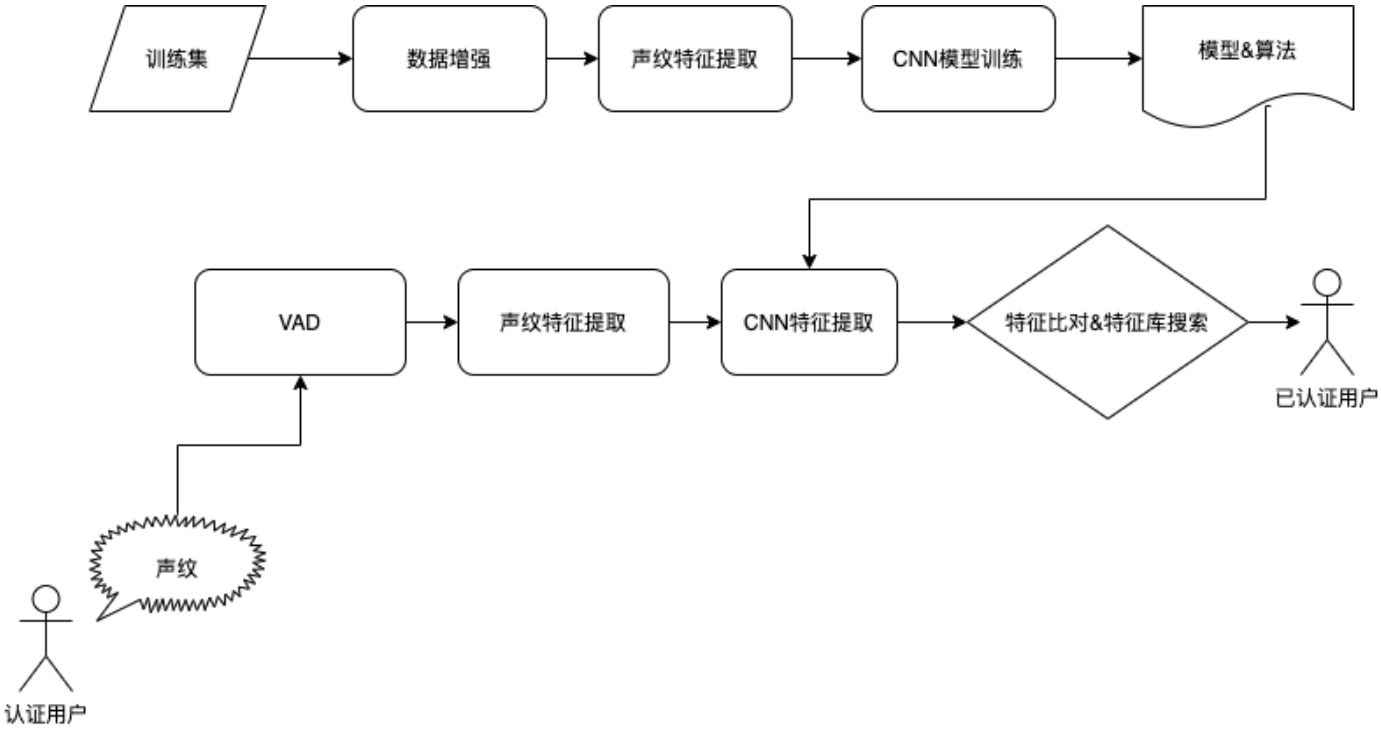
AnalyticDB向量检索+AI 实战: 声纹识别

1. 背景

近年来，随着人工智能对传统行业的赋能改造，越来越多的基于人工智能的业务解决方案被提出来，声纹识别在保险行业中的身份认证便是一个很好的例子. 声纹识别是根据说话人发音的生理和行为特征，自动识别说话人身份的一种生物识别技术，对应在电话销售场景下，它主要解决以下安全问题：一方面，有不法分子窃取电话销售人员账号信息，非法获取客户个人信息资料并进行贩卖、泄露，严重侵犯了公民个人的信息隐私权，另一方面，部分行业从业人员利用一些规则漏洞，通过套保、骗保等非法手段实施金融诈骗. 针对这些安全问题，可以通过实时声纹认证加以解决，以电话销售人员为监管核心，利用每个人独一无二的声纹进行严密的个人身份认证，保证电话销售人员对接客户时是本人注册登录，规范电销人员行为，从源头上有效规避信息泄露、漏洞利用等风险.

2. 声纹识别原理

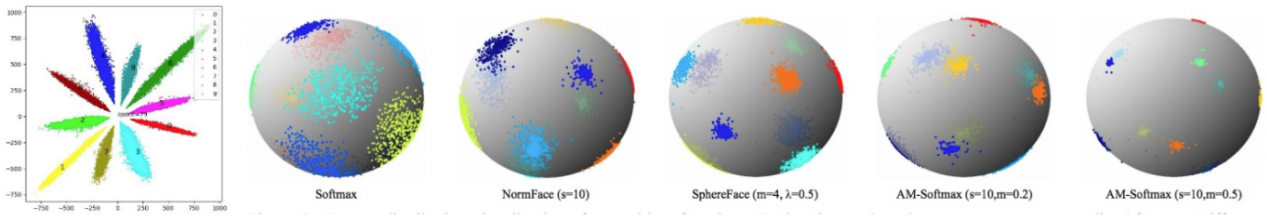


上图是端对端的深度学习训练和推理过程。对比传统声纹识别模型，我们的模型在实际使用中优势明显，在用户远程身份验证场景，通过注册用户说一段话，即可轻松快速的确认注册用户身份，识别准确率达到95%以上，秒级响应，实时声纹核身。下面简要介绍我们模型的特点.

2.0 度量学习

实验发现，在声纹识别中采用softmax进行网络训练 ，用余弦相似度的测试性能往往不如传统声纹识别模型，尤其是在鲁棒性上。分析发现[6]基于softmax的分类训练，为了得到更小的loss，优化器会增大一些easy samples的L2 length，减小hard examples 的L2 length，导致这些样本并没有充分学习，特征呈现放射状，以MNIST识别任务为例，基于softmax学到的特征分布如图3(a)所示. 同类别特征分布并不聚拢，在L2 长度上拉长，呈放射状，且每个类别的间距并不大，在verification的任务中，会导致相邻的两个类别得分很高.

为了达到类内聚拢，类间分散的效果，我们研究了在图像领域中应用较为成功的几种softmax变种，包括AM-softmax[4]，arcsoftmax[5]等，从图3(b)可以看到，基于margin的softmax，相比纯softmax，类间的分散程度更大，且类内特征更聚拢，对声纹1:1比对和1:N搜索的任务友好.



(a) (b)

图3. (a) MNIST数据集在softmax训练策略下特征二维化后的分布 （b）不同训练策略下特征长度归一化后的分布

2.1 噪音鲁棒性

在特征提取时，对于简单加性噪音，我们提出了基于功率谱减法，实现噪音抑制；对于其他复杂噪音，我们提出了基于降噪自动编码器的噪音补偿模型，将带噪语音特征映射到干净语音特征，实现噪音消除.
在模型训练时，我们采用数据增强的训练机制，将噪音数据通过随机高斯的形式加入到声纹模型的训练中，使得训练后的模型对噪音数据具有更好的鲁棒性.

2.2 短音频鲁棒性

为了提高短音频鲁棒性，我们提出了基于短时帧级别的模型训练机制，使模型能够在极短的语音时长（约0.5秒）下即可完成声纹识别. 在此基础上，我们在模型训练中引入了更多高阶的音频统计信息和正则化方法，进一步提升了模型在短语音条件下（2~3秒）的识别精度.

3. 如何使用AnalyticDB搭建声纹比对系统

3.0 创建插件

使用一下SQL来分别创建AnalyticDB的非结构化分析插件OpenAnalytic和向量检索插件fastann.

```
CREATE EXTENSION IF NOT EXISTS open_analytic;
CREATE EXTENSION IF NOT EXISTS fastann;
```

3.1 建表

我们可以建立一个表来保存所有说话人的声音和声音的特征, 后续我们可以从这个表中搜索说话人.

```
CREATE TABLE speaker_table (
        speaker_id TEXT NOT NULL,  # 说话人id
        audio BYTEA NOT NULL, # 声音文件
        feature REAL[] NOT NULL, # 声音特征向量
        PRIMARY KEY (question)
);
```

3.2 创建索引

我们可以为特征向量列创建向量检索索引.

```
CREATE INDEX speaker_table_index
ON speaker_table USING ann (feature) WITH (dim = 512);
```

3.3 创建声纹识别算法pipeline

通过以下sql, 我们可以在数据库中创建声纹特征提取的算法模型.

```
SELECT open_analytic.pipeline_create('speaker_feature_extractor');
```

3.4 获取说话人声纹特征

通过以下sql可以使用3.3创建的pipeline. 这个UDF的输入是pipeline名称和目标文本. 输出是一个说话人声音的特征向量.

```
# 通过声音文件识别
SELECT open_analytic.pipeline_run_dist_random('speaker_feature_extractor',
                                                <声音文件>);
```

3.5 说话人声纹特征导入AnalyticDB

获取声音特征后, 我们可以使用一下sql来讲数据插入3.1创建的表中.

```
INSERT INTO speaker_table VALUES (<说话人id>,  <声音文件>, <声音特征向量>);
```

3.6 在数据库中搜索最相似的的人

通过以下sql, 我们可以在声音特征库中搜索最相似的说话人. 然后我们可以根据特征间距离是否满足预设的阈值来判断是否是同一个人.

```
SELECT speaker_id, l2_distance(feature,  <声音特征向量>)
```

```
ORDER BY feature <-> <声音特征向量>
LIMIT 10;
```

3.7 比较两个声音是否为同一个人

我们还可以提取出两个人的声音特征然后直接计算二者的距离来判断这两个声音是否来自同一个说话人. SQL如下

```
SELECT  l2_distance(feature1,  feature2);
```

4. AnalyticDB介绍

分析型数据库(AnalyticDB)是阿里云上的一种高并发低延时的PB级实时数据仓库，可以毫秒级针对万亿级数据进行即时的多维分析透视和业务探索。AnalyticDB for MySQL 全面兼容MySQL协议以及SQL:2003语法标准, AnalyticDB forPostgreSQL 支持标准 SQL:2003，高度兼容 Oracle 语法生态.
向量检索和非结构化数据分析是AnalyticDB的进阶功能. 目前两款产品都包含向量检索功能, 可以支持人脸, 人体, 车辆等的相似查询和推荐系统。AnalyticDB在真实应用场景中可以支持10亿级别的向量数据的查询, 毫秒级别的响应时间. AnalyticDB已经在多个城市的重大项目中大规模部署.
在一般的包含向量检索的的应用系统中, 通常开发者会使用向量检索引擎(例如Faiss)来存储向量数据, 然后使用关系型数据库存储结构化数据. 在查询时也需要交替查询两个系统, 这种方案会有额外的开发工作并且性能也不是最优. AnalyticDB支持结构化数据和非结构化数据(向量)的检索，仅仅使用SQL接口就可以快速的搭建起以图搜图或者图片+结构化数据混合检索等功能. AnalyticDB的优化器在混合检索场景中会根据数据的分布和查询的条件选择最优的执行计划，在保证召回的同时，得到最优的性能。AnalyticDB向量版采用了多项创新性技术, 这些技术在我们的论文 *AnalyticDB-V: A Hybrid Analytical Engine Towards Query Fusion for Structured and Unstructured Data* 中有详细介绍介绍. 目前论文已经被数据库三大顶会之一的VLDB接受, 具有技术领先性.
结构化信息+非结构化信息(图片)混合检索在实际应用中被广泛使用的. 例如人脸门禁系统被部署在多个小区时, 我们使用一张表存储了所有小区的人脸特征, 在人脸检索时我们只需要检索当前小区的人脸特征. 在这种情况下, 使用AnalyticDB我们只需要在SQL中增加where 小区名 ='xxx' 就可以轻易实现。
AnalyticDB同时提供了先进的图像文本分析算法, 能够提取非结构化数据的特征和标签, 用户仅仅需要使用SQL就可以完成图像文本内容的分析.
更多信息可以参考文章: https://zhuanlan.zhihu.com/p/82284704

5. 参考文献

[1] Heigold G, Moreno I, Bengio S, et al. End-to-end text-dependent speaker verification[C]//2016 IEEE International Conference on Acoustics, Speech and Signal

Processing (ICASSP). IEEE, 2016: 5115-5119.
[2]Li C, Ma X, Jiang B, et al. Deep speaker: an end-to-end neural speaker embedding system[J]. arXiv preprint arXiv:1705.02304, 2017.
[3] Snyder D, Garcia-Romero D, Sell G, et al. X-vectors: Robust den embeddings for speaker recognition[C]//2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2018: 5329-5333.
[4] Wang F, Cheng J, Liu W, et al. Additive margin softmax for face verification[J]. IEEE Signal Processing Letters, 2018, 25(7): 926-930.
[5] Dang J, Guo J, Xue N, et al. Arc face: Additive angular margin loss for deep face recognition[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2019: 4690-4699.
[6] Ranjan R, Castillo C D, Chellappa R. L2-constrained softmax loss for discriminative face verification[J]. arXiv preprint arXiv:1703.09507, 2017.

6. 结语

本文介绍了如何使用AnalyticDB来搭建声纹比对系统. AnalyticDB还支持其他多种多样人工智能算法如目标检测, 商品识别, 基因识别等等. 想了解更多请扫码加入AnalyticDB向量版交流群.

扫一扫群二维码，立刻加入该群。

7. 往期文献

[1] 戴口罩也能刷门禁？疫情下AnalyticDB亮出社区管理的宝藏神器!
https://developer.aliyun.com/article/745160
[2] 阿里云提供高效基因序列检索功能，助力冠状病毒序列快速分析
https://developer.aliyun.com/article/753097
[3] 三步搭建一套声纹系统

https://developer.aliyun.com/article/765232
[4] 阿里云提供高效病原体检测工具助力精准医疗
https://yq.aliyun.com/articles/761891
[5] 使用AnalyticDB轻松实现以图搜图和人脸检索
https://developer.aliyun.com/article/765982
[6] 三步在阿里云上面搭建一套个性化推荐系统
https://blog.csdn.net/yunqiinsight/article/details/107166815