Triplet loss functions in speaker recognition systems

Francisco Javier Sáez Maldonado

May 10, 2022

Máster en Ciencia de Datos

Escuela Politécnica Superior
Universidad Autónoma de Madrid

Introduction

- **Task**: Speaker Recognition
  - 'Closed-set' vs 'Open-set'

Introduction

- **Task**: Speaker Recognition
    - 'Closed-set' vs 'Open-set'
- **Data**: VoxCeleb challenge dataset.

Original Paper

In defence of metric learning for speaker recognition

Joon Son Chung, Jaesung Huh, Seongkyu Mun, Minjae Lee, Hee Soo Heo, Soyeon Choe, Chiheon Ham, Sunghwan Jung, Bong-Jin Lee, Icksang Han

Naver Corporation, South Korea

joonson.chung@navercorp.com

Abstract

The objective of this paper is 'open-set' speaker recognition of unseen speakers, where ideal embeddings should be able to condense information into a compact utterance-level representation that has small intra-speaker and large inter-speaker distance.

A popular belief in speaker recognition is that networks trained with classification objectives outperform metric learning methods. In this paper, we present an extensive evaluation of most popular loss functions for speaker recognition on the VoxCeleb dataset. We demonstrate that the vanilla triplet loss shows competitive performance compared to classification-based losses, and those trained with our proposed metric learning objective outperform state-of-the-art methods.

popular due to their ease of implementation and good performance [17, 18, 19, 20, 21, 22, 23, 24]. However, training with AM-Softmax and AAM-Softmax has proven to be challenging since they are sensitive to the value of scale and margin in the loss function.

Metric learning objectives present strong alternatives to the prevailing classification-based methods, by learning embeddings directly. Since open-set speaker recognition is essentially a metric learning problem, the key is to learn features that have small intra-class and large inter-class distance. Contrastive loss [25] and triplet loss [26] have been demonstrated promising performance on speaker recognition [27, 28] by optimising the distance metrics directly, but these methods require careful pair or triplet selection which can be time consuming and performance sensitive.

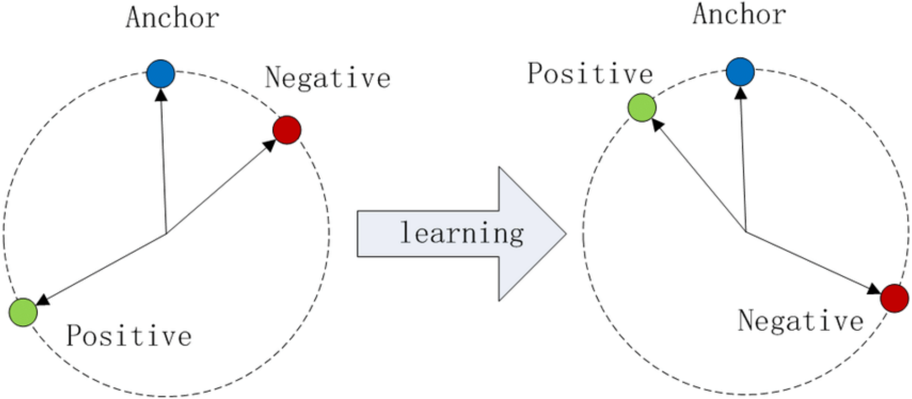
Figure 1: Original paper Chung u. a. (2020).

Softmax is not enough

- **Softmax:** $L_S = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{e^{\mathbf{W}_{y_i}^T x_i + b_{y_i}}}{\sum_{j=1}^{C} e^{\mathbf{W}_j^T x_i + b_j}}$

Softmax is not enough

- **Softmax:** $L_S = -\frac{1}{N}\sum_{i=1}^{N} \log \frac{e^{\mathbf{W}_{y_i}^T x_i + b_{y_i}}}{\sum_{j=1}^{C} e^{\mathbf{W}_j^T x_i + b_j}}$



Triplet Loss

$$\left\|g(x) - g(x^{+})\right\|_2 + \alpha < \left\|g(x) - g(x^{-})\right\|_2.$$

Triplet Loss

$$\left\|g(x) - g(x^+)\right\|_2 + \alpha < \left\|g(x) - g(x^-)\right\|_2 .$$

Definition (Triplet loss term)

Given an anchor x, a positive sample x^+ and a negative sample x^-, a term of the triplet loss function is defined as:

$$\ell^\alpha(x, x^+, x^-) = \max\left(0, \left\|g(x) - g(x^+)\right\|_2^2 - \left\|g(x) - g(x^-)\right\|_2^2 + \alpha\right) . \tag{1}$$

$$\mathcal{L}(x_i, x_i^+, x_i^-) = \sum_{i \in \Lambda} \ell^\alpha(x_i, x_i^+, x_i^-). \tag{2}$$

Searching for negative samples - Hard negative mining

Firstly: we need two utterances from each speaker: anchor x and positive x^+.

Algorithm 1 Hard negative mining

```
1: for Each audio in the batch do
2:     Take x and x⁺.
3:     Compute squared pairwise distance between x and x⁺.
4:     Use computed distances to extract hard negative.
5: end for
```

Results

Model	Encoder	Loss function	EER
Pretrained	ResNetSE34L	Angle Proto	2.1792
	ResNetSE34V2	Softmax Proto	1.1771
Assignment	ResNetSE34L	Amsofmax	17.60
This presentation	ResNetSE34L	Triplet	20.73

Table 1: Execution results.

Possible improvements

Problem: We are only comparing to one example in each iteration!

Possible improvements

Problem: We are only comparing to one example in each iteration!

Definition

Let x^+ *be a positive example of the anchor* x*, and consider the set* $X^- = \{x_1^-, \cdots, x_{N-1}^-\}$ *of* $(N-1)$ *negative samples. Given an encoder* g*, the* $(N+1)-$*tuplet loss is defined as follows:*

$$\mathcal{L}_{(N+1)-tuplet}(x, x^+, X^-) = \log\left(1 + \sum_{i=1}^{N-1} \exp\left(g(x)^T g(x_i^-) - g(x)^T g(x^+)\right)\right) \quad (3)$$

Conclusions

- There is a couple of loss functions that can be used in the speaker recognition task.

Conclusions

- There is a couple of loss functions that can be used in the speaker recognition task.
- Triplet losses are competitive and intuitive.

Conclusions

- There is a couple of loss functions that can be used in the speaker recognition task.
- Triplet losses are competitive and intuitive.
- There is room for improvement.

Thank you for your attention.

References

[Chung u. a. 2020] Chung, Joon S. ; Huh, Jaesung ; Mun, Seongkyu ; Lee, Minjae ; Heo, Hee-Soo ; Choe, Soyeon ; Ham, Chiheon ; Jung, Sung-Ye ; Lee, Bong-Jin ; Han, Icksang: In defence of metric learning for speaker recognition. In: *ArXiv* abs/2003.11982 (2020)