Centroid-based Deep Metric Learning For Speaker Recognition

Jixuan Wang[1,2,*], Kuan-Chieh Wang[1,2,*], Marc T. Law[1,2], Frank Rudzicz[1,2,3,4], Michael Brudno[1,2,5]

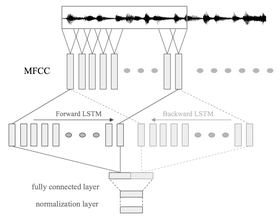
[1]University of Toronto, Canada [2]Vector Institute, Canada [3]St Michaels Hospital, Canada
[4]Surgical Safety Technologies Inc, Canada [5]Hospital for Sick Children, Canada

Introduction

- We propose the use of **prototypical network loss** (PNL), a state-of-the-art approach for the few-shot image classification task, to optimize an end-to-end speaker embedding network.
- We showed that models trained using PNL outperform models of the same architecture optimized with **triplet loss** (TL) on speaker verification (SV) and speaker identification (SI) tasks.

Optimization schemes & Model

- $\mathbf{x}$: a sequence of speech features
- y: speaker label
- d: distance function, e.g. cosine or squared Euclidean distance
- f: speech sequence embedding model



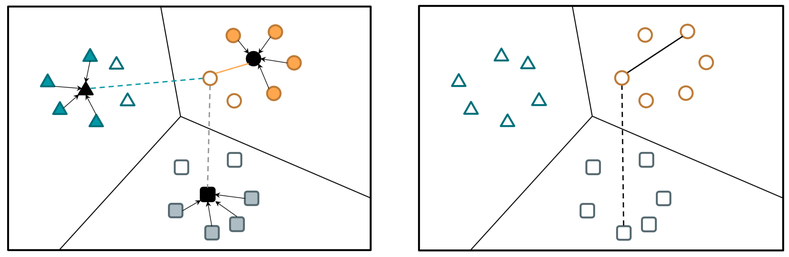
Prototypical Networks [1] train a neural network episodically. Each mini-batch contains a support set, $S = \{(\mathbf{x}_i, y_i)\}_{i=1}^{N_S}$, and a query set, $Q = \{(\mathbf{x}_j, y_j)\}_{j=1}^{N_Q}$. A query $\mathbf{x}_j$ is classified based on how close it is to the class *prototype* $\mathbf{c}_{y_j}$ of class y_j (computed as the average of $f(\mathbf{x})$ for all $\mathbf{x}$ in the support set S_{y_j} of class y_j): $p(y = y_j|\mathbf{x}_j) = \exp\left(-d(f(\mathbf{x}_j), \mathbf{c}_{y_j})\right) / \Sigma_{k'} \exp(-d(f(\mathbf{x}_j), \mathbf{c}_{k'})$. The loss function for each mini-batch is:

$$J_{PNL} = \sum_{\{(\mathbf{x}_j, y_j)\} \in Q} -\log p(y = y_j|\mathbf{x}_j) \quad (1)$$

Triplet-based models [2] sample triplets, which consist of an anchor $\mathbf{x}_a$, a positive sample $\mathbf{x}_p$ with the same speaker label, and a negative sample $\mathbf{x}_n$ with a different speaker label. By sampling all possible triplets in a mini-batch, the loss for this mini-batch is:

$$J_{TL} = \sum_{\tau \in T} max(0, d(f(\mathbf{x}_a^\tau), f(\mathbf{x}_p^\tau)) - d(f(\mathbf{x}_a^\tau), f(\mathbf{x}_n^\tau)) + \alpha) \quad (2)$$

Prototypical Network Loss vs. Triplet Loss



(a) prototypical network loss (b) triplet loss

Dashed lines represent distances encouraged to increase, while **solid lines** represent distances being decreased.

- **PNL**: prototypes for different speakers, denoted by black nodes are computed as the mean of the support set (shaded) during training.
- **TL**: a triplet consists of an anchor, positive, and negative samples, forming the (anchor, positive) and (anchor,negative) pairs. Depending on the sampling strategy not all triplets may be considered.

Conclusion

- With identical speech sequence embedding architectures, PNL **outperforms** the triplet loss when speakers are seen during training, and by **an even larger margin** on held-out, unseen speakers for both speaker identification and speaker verification tasks.
- In the future, we would like to explore better architectures of speech sequence embedding models, compare PNL with other loss functions used in deep metric learning and integrate the proposed model into a speaker diarization pipeline.

References

[1] Snell, Jake, Kevin Swersky, and Richard Zemel. "Prototypical networks for few-shot learning." *Advances in Neural Information Processing Systems*. 2017.
[2] Bredin, Hervé. "Tristounet: triplet loss for speaker turn embedding." 2017 *IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2017.

Results on Speaker Identification

SI accuracy on test and unseen sets of VCTK and VoxCeleb2. Under the "Model" column are the training configurations. The top row '$S{:}N_S$, $Q{:}N_Q$, K-way' denotes the *task* configurations.

(a) SI accuracy on VCTK

Model	S: 5, Q: 5, 6-way		S: 10, Q: 10, 18-way	
	Test	Unseen	Test	Unseen
TL (Naive, Euc)	91.96%	77.80%	81.25%	56.37%
TL (Naive, Cos)	92.37%	77.69%	83.51%	58.51%
TL (Semi, Euc)	93.33%	79.69%	85.38%	58.49%
TL (Semi, Cos)	92.13%	73.94%	83.99%	52.97%
PNL (1_5, Euc)	93.24%	77.90%	83.71%	56.52%
PNL (3_5, Euc)	**95.63%**	84.81%	**90.53%**	**69.64%**
PNL (5_5, Euc)	95.47%	83.69%	89.85%	68.55%
PNL (10_10, Euc)	94.38%	**85.00%**	88.43%	66.63%

(b) SI accuracy on VoxCeleb2

Model	S: 10, Q: -, 15-way		S: 30, Q: -, 15-way	
	Test	Unseen	Test	Unseen
TL (Semi, Cos)	74.74%	53.92%	75.18%	59.61%
TL (Semi, Euc)	71.78%	51.74%	72.02%	56.79%
PNL (5_5, Euc)	**78.38%**	**59.44%**	**79.23%**	**66.63%**

Results on Speaker Verification

EER of SV on both VCTK and VoxCeleb2 datasets. "60s" (60 seconds) refers the duration of speech we used for enrollment.

(a) EER on VCTK

Model	Test	Unseen	
	60s	60s	10s
TL (Semi, Cos)	5.43(±0.16)	13.87(±0.37)	16.19(±0.86)
TL (Semi, Euc)	5.05(±0.09)	12.26(±0.69)	13.44(±0.91)
PNL (5_5, Euc)	**4.08(±0.13)**	**10.77(±0.58)**	**12.00(±0.76)**

(b) EER on VoxCeleb2

Model	Test	Unseen	
	60s	60s	10s
TL (Semi, Cos)	9.23(±0.13)	14.62(±0.35)	16.93(±0.45)
TL (Semi, Euc)	9.90(±0.11)	15.92(±0.32)	17.61(±0.51)
PNL (5_5, Euc)	**8.29(±0.12)**	**13.68(±0.26)**	**15.67(±0.56)**