Continuous Metric Learning For Transferable Speech Emotion Recognition and Embedding Across Low-resource Languages

Sneha Das[1], Nicklas Leander Lund[1], Nicole Nadine Lønfeldt[2], Anne Katrine Pagsberg[2, 3], Line H. Clemmensen[1]

Motivation

- Speech Emotion Recognition (SER): Infer emotional state of individuals from speech signals.
- SER applications: Commercial sector, education, healthcare.
- Challenges: Generalization over languages, corpora; interpretability.

Objective and Proposal

- Goal: Obtain emotion representations from speech that are transferable to low-resource (data and labels) languages.
- Proposal: Semi-supervised DAE→to shape the latent space with emotion-relevant information.
- Contributions:
  - Method for continuous metric learning to order samples in latent space.
  - Data labels with activation and valence annotations for open datasets.

Methodology

- Semi-supervision using activation and valence labels.
- DAE Loss function: Minimize reconstruction loss + similar distance between embedding and labels.
- Trained on single (large) dataset, tested on transfer (>4) datasets.

$$\arg\min_{f_\theta, g_\phi} \quad \mathcal{L}_{\text{rec}} + \mathcal{L}_{\text{met}} = \mathcal{L}_{\text{rec}} + \mathcal{L}_{\text{res}} + \mathcal{L}_{\text{sl}}$$

$$\mathcal{L}_{\text{res}} = \mathbb{E}\|\mathbf{z_d} - \hat{\mathbf{z}}_\mathbf{d}\|_2^2, \quad \hat{z}_d = pl_d,$$

$$p = ({\mathbf{l_d}}^T\mathbf{l_d})^{-1}{\mathbf{l_d}}^T\mathbf{z_d}$$

$$\mathcal{L}_{\text{sl}} = \left\| \frac{\hat{z}_d(a) - \hat{z}_d(b)}{l_d(a) - l_d(b)} - 1 \right\|_2.$$

Table 1: Adjusted squared correlation coefficient presenting the linear dependence of z_d on l_d for the three models. Mean and standard deviation over five folds are presented.

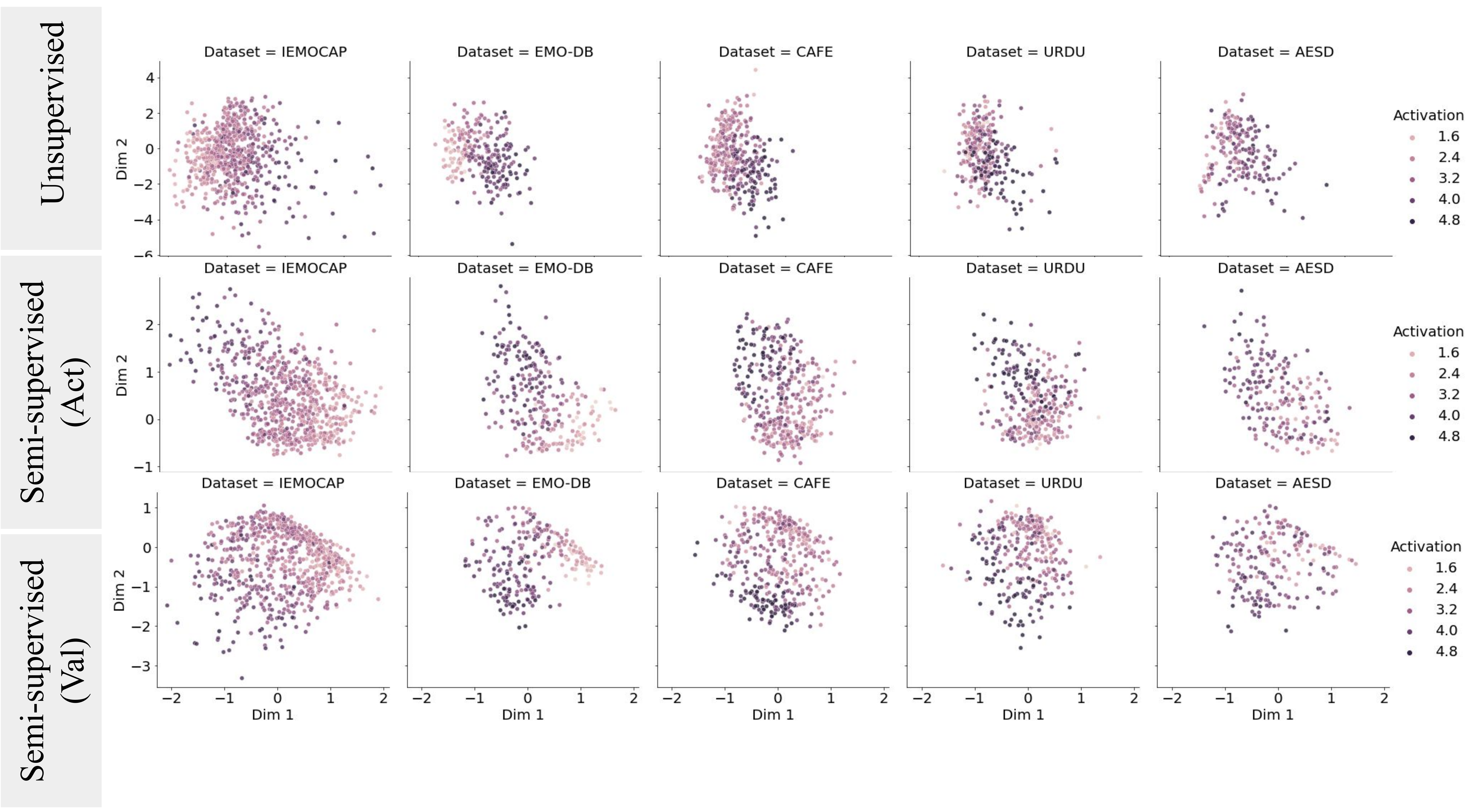
Method	R^2-Act ($\mu \pm \sigma$)	R^2-Val ($\mu \pm \sigma$)
Unsupervised	0.11±0.06	0.03±0.02
Metric-act	**0.21 ± 0.05**	0.06±0.02
Metric-val	0.12± 0.05	0.05±0.02

Table 2: Adjusted squared correlation coefficient presenting the linear dependence of l on z, the activation and valence labels for the three models. Mean and standard deviation over five folds are presented.

Method	IEMOCAP		EMO-DB		CAFE		URDU		AESD	
	R^2-Act ($\mu \pm \sigma$)	R^2-Val ($\mu \pm \sigma$)	R^2-Act ($\mu \pm \sigma$)	R^2-Val ($\mu \pm \sigma$)	R^2-Act ($\mu \pm \sigma$)	R^2-Val ($\mu \pm \sigma$)	R^2-Act ($\mu \pm \sigma$)	R^2-Val ($\mu \pm \sigma$)	R^2-Act ($\mu \pm \sigma$)	R^2-Val ($\mu \pm \sigma$)
Unsupervised	0.41 ± 0.03	0.05 ± 0.02	0.62 ± 0.04	0.06 ± 0.04	0.41 ± 0.03	0.14 ± 0.02	0.28 ± 0.05	0.14 ± 0.02	0.3 ± 0.01	$0.0 \pm 0.0^*$
Metric-act	$\mathbf{0.49 \pm 0.02}$	0.05 ± 0.01	$\mathbf{0.63 \pm 0.02}$	0.04 ± 0.03	0.45 ± 0.03	0.13 ± 0.03	0.33 ± 0.04	0.14 ± 0.03	$\mathbf{0.31 \pm 0.05}$	$-0.0 \pm 0.0^*$
Metric-val	0.4 ± 0.03	$\mathbf{0.11 \pm 0.01}$	0.6 ± 0.03	$\mathbf{0.1 \pm 0.0}$	$\mathbf{0.45 \pm 0.01}$	$\mathbf{0.14 \pm 0.03}$	$\mathbf{0.37 \pm 0.03}$	$\mathbf{0.17 \pm 0.02}$	0.29 ± 0.02	$0.01 \pm 0.01^*$

Conclusions

- Proposed semi supervision method yields more consistent latent space.
- Consistent latent space →transferable to unseen (low-resource) languages.



NLDL 2022

DTU

REGION H

Psykiatri

1. Department of Applied Mathematics and Computer Science, Technical University of Denmark
2. Child and Adolescent Mental Health Center, Copenhagen University Hospital, Capital Region, Copenhagen
3. Faculty of Health, Department of Clinical Medicine, Copenhagen University