Towards Transferable Speech Emotion Representation: On Loss Functions For Cross-Lingual Latent Representations

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

[1]Department of Applied Mathematics and Computer Science, Technical University of Denmark

[2]Child and Adolescent Mental Health Center, Copenhagen University Hospital, Capital Region

[3]Faculty of Health, Department of Clinical Medicine, Copenhagen University

DTU Compute
Department of Applied Mathematics and Computer Science

Motivation

- Speech emotion recognition (SER): inferring emotional state from speech signals.
- Emotion recognition employed in healthcare, education sector, criminal justice system.
- SER: signal processing, machine learning, deep learning.
- Existing challenges: Generalizing over languages, corpora, recording condition (under low-resource conditions).

Objectives and Contributions

Objectives for transferability:

1. Latent embedding with discrimination between emotion classes.
2. Latent distribution that are consistent over corpora.

Contributions:

1. Low-complexity DAE and VAE.
2. VAE with KL-loss annealing: balancing KL-loss and reconstruction loss.
3. VAE with semi-supervision incorporating clustering in latent space.

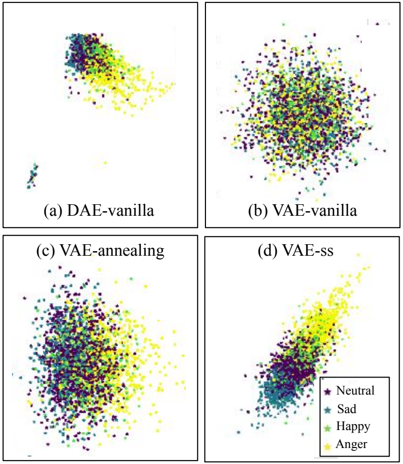
Formulation

- DAE:

$$\underset{f_\theta, g_\phi}{\arg\min} \quad \mathcal{L}_{\text{rec}} = \mathbb{E}\|\mathbf{x} - g_\phi(f_\theta(\mathbf{x_n}))\|_2^2, \tag{1}$$

- VAE:

$$\underset{\theta, \phi}{\arg\min} \quad \mathcal{L}_{\text{rec}} + \mathcal{L}_{\text{KL}} = -\mathbb{E}_{\mathbf{z}\sim q_\theta(\mathbf{z}|\mathbf{x})} \log p_\phi(\mathbf{x}|\mathbf{z}) + D_{KL}(q_\theta(\mathbf{z}|\mathbf{x})||p(\mathbf{z})), \tag{2}$$



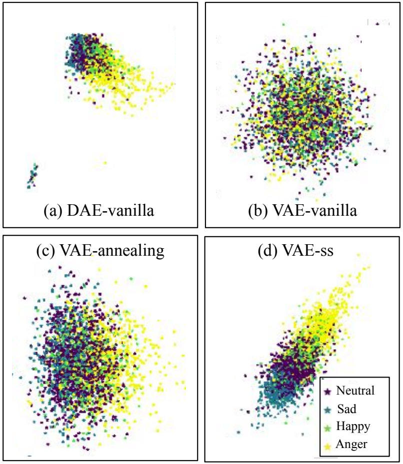
Formulation

- VAE with KL-annealing:

$$\arg\min_{\theta,\phi} \quad \mathcal{L}_{\text{rec}} + \mathcal{L}_{\text{KL}} = -\mathbb{E}_{\mathbf{z}\sim q_\theta(\mathbf{z}|\mathbf{x})} \log p_\phi(\mathbf{x}|\mathbf{z}) \\ +\beta_e D_{KL}(q_\theta(\mathbf{z}|\mathbf{x})||p(\mathbf{z})), \tag{3}$$

where the standard formulation of β_e:

$$\beta_e = \begin{cases} f(\tau) = \frac{0.25}{R}\tau, & \tau \leq R \\ 0.25, & \tau > R \end{cases} \quad \text{where} \quad \tau = \frac{\text{mod}(e-1, \frac{T}{M})}{\frac{T}{M}}, \tag{4}$$

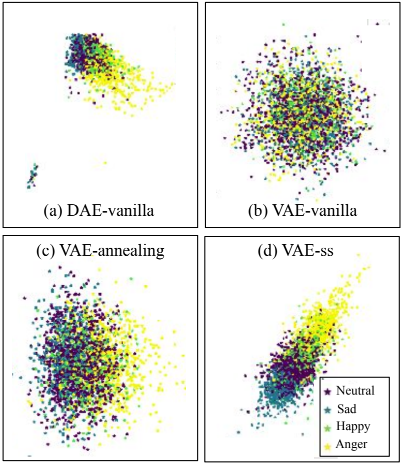


Formulation

- VAE with semi-supervision:

$$\underset{\theta,\phi}{\arg\min} \quad \mathcal{L}_{\text{rec}} + \beta_e \mathcal{L}_{\text{KL}} + \gamma \mathcal{L}_{\text{clus}},$$

$$\mathcal{L}_{\text{clus}} = \frac{D_{\text{intra}}}{D_{\text{inter}}} = \frac{\sum_{k=1}^{K} \sum_{\forall i \in k} D(\mathbf{z_i}, \overline{\mathbf{z}}^{\mathbf{k}})}{\sum_{k=1}^{K-1} \sum_{j=k+1}^{K} D(\overline{\mathbf{z}}^{\mathbf{k}}, \overline{\mathbf{z}}^{\mathbf{j}})}, \tag{5}$$



Architecture

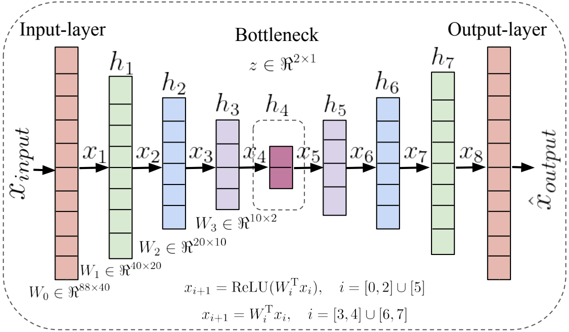


Figure: Illustration of the architecture employed for all the models explored in this work.

- Training: 50 epochs, batch size 64, Adam optimizer (learning rate: 1e-3).
- Latent embedding used as input features to a linear SVC.

Evaluation

- Datasets: IEMOCAP, SAVEE, Emo-DB, CaFE, URDU, AESD
- Input features: eGeMAPS using OpenSmile
- Preprocessing: remove outliers using z-score normalization $(-10 > z > 10)$
- 5-fold cross validation

Results: Classification performance

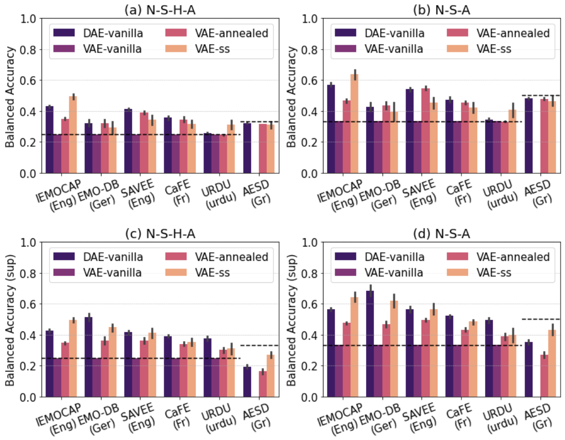


Figure: (1) Balanced accuracy on unseen transfer data sets using (a) 4 emotion classes, (b) 3 emotion classes; balanced accuracy with access to 20% of the unlabeled transfer data sets with (c) 4 emotions and (d) 3 emotion classes.

Results: Consistency of latent space

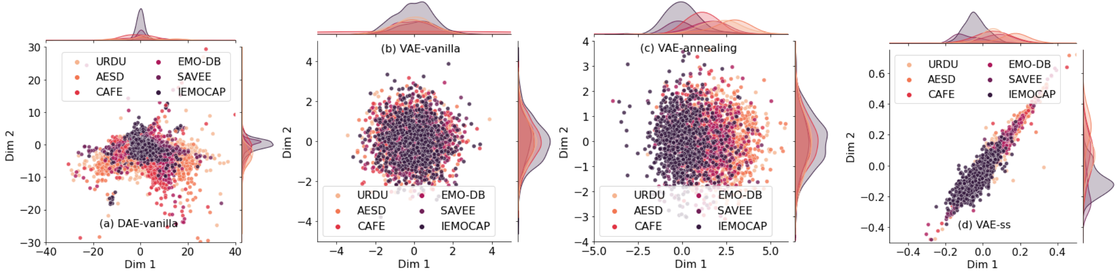
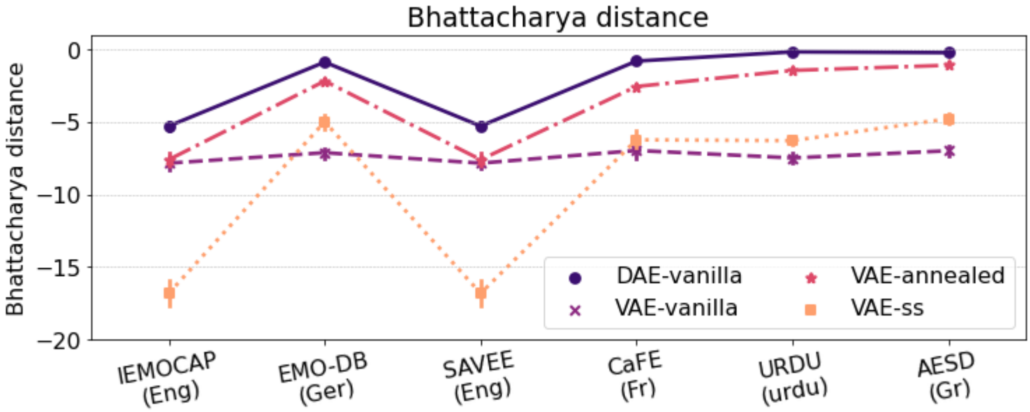


Figure: Scatter plots depicting the overlap between the latent embedding obtained from the methods investigated for all the transfer data sets.

Results: Consistency of latent space



Conclusions

1. DAE: highest classification accuracy, worst distribution consistency.
2. VAE-vanilla: best consistency, classification accuracy random.
3. VAE-ss: Classification accuracy similar to DAE and distribution consistency similar to VAE-vanilla.