The Bridge from D to E: Exploring the alignment of generative latent spaces with the VA space

Surabhi S Nath

2016271

Vishaal Udandarao

2016119

Motivation

- Emotions are popularly represented through Valence (V) and Arousal (A)
- Generating a latent space highly aligned with VA will enable a more descriptive and disentangled representation of the emotional images
- This can serve several applications such as:
  - Improving robustness of emotional classifiers
  - Data augmentation
  - Facial expression editing

Objective

To model the VA space using a latent generative model like VAEs

- Given a dataset of emotional face images, with VA annotation, measure the alignment/divergence of latent space with VA values, without explicit supervision
- Repeat the experiment with a regularization loss term to enforce the VAE to mimic the VA emotional space
- Compare performance of the 2 models for 2 datasets
- For datasets without VA annotations, obtain VA approximates using transfer across datasets

Related Work

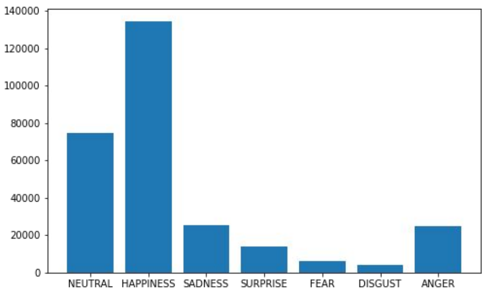
- Latif et al. [1] used a VAE + LSTM hybrid to classify emotional speech.
- Marmpena et al. [2] utilised VAEs for generating emotional body language animations for a robot.
- Suguitan et al. [3] proposed a method for modifying affective robot movements using VAEs.
- Lindt et al. [4] used a generative model for automated facial expression editing along continuous valence and arousal dimensions.
- Kollias et al. [5] developed a Morphable Model which modified the neutral image reconstruction by adding affect followed by blending with the original image.

Methodology

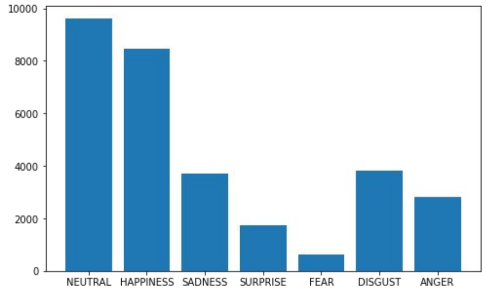
Datasets

1. **Affectnet**: ~420K annotated images with continuous VA values in [-1, 1] along with emotional labels: Neutral, Anger, Happiness, Sadness, Surprise, Fear, Disgust, Contempt, None, Uncertain, Non-face
2. **IMFDB**: ~34K annotated images of 100 Indian actors, no VA supervision, emotional labels: Neutral, Anger, Happiness, Sadness, Surprise, Fear, Disgust
3. **AFEW**: ~24K annotated images from videos of ~600 actors with only discrete VA values in range [-10, 10]

Class Distributions



Affectnet



IMFDB

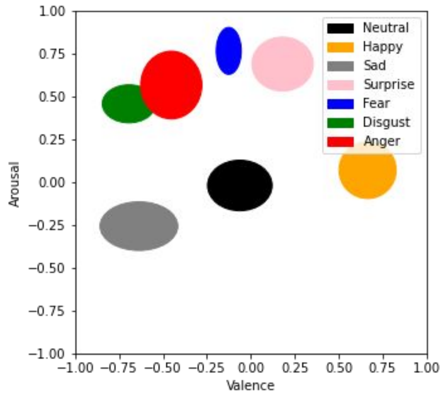
Methodology

Annotation Transfer

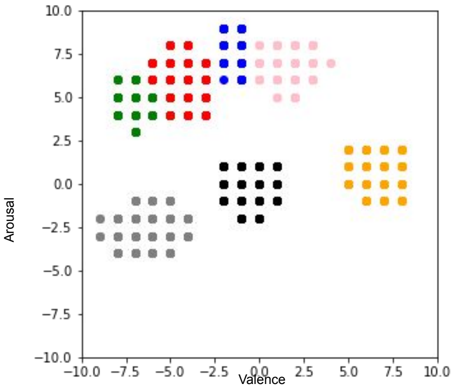
- While IMFDB did not have VA annotations, Affectnet had continuous VA labels and AFEW had discrete VA labels
- Consider Affectnet as anchor dataset due to its large size and similar domain as IMFDB
- Continuous and Discrete VA labels for IMFDB were obtained using the **ellipse sampling method**
- Ellipses for each label were obtained using Affectnet with centre as **(mean valence value, mean arousal value)**, semi major axis as the **standard deviation of valence values** and semi minor axis as the **standard deviation of arousal values**

Methodology

Annotation Transfer

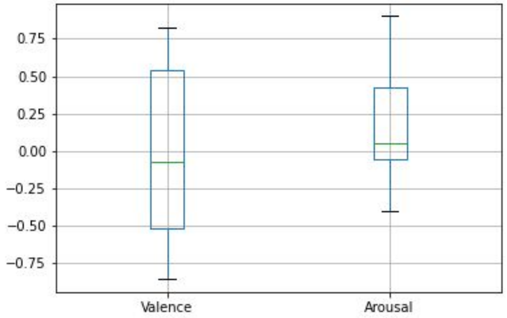


- A value was sampled randomly with equal likelihood from the ellipse corresponding to the IMFDB image label, from which VA value was achieved.
- For discrete labels, the Affectnet VA values were scaled from -10 to 10 and sampling of only integral values was performed from the corresponding scaled ellipses.

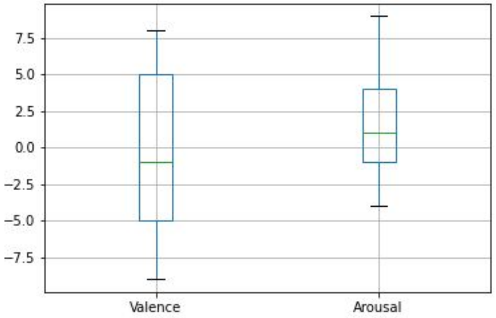


Methodology

Transferred Annotations



Transferred continuous VA annotations (IMFDB)

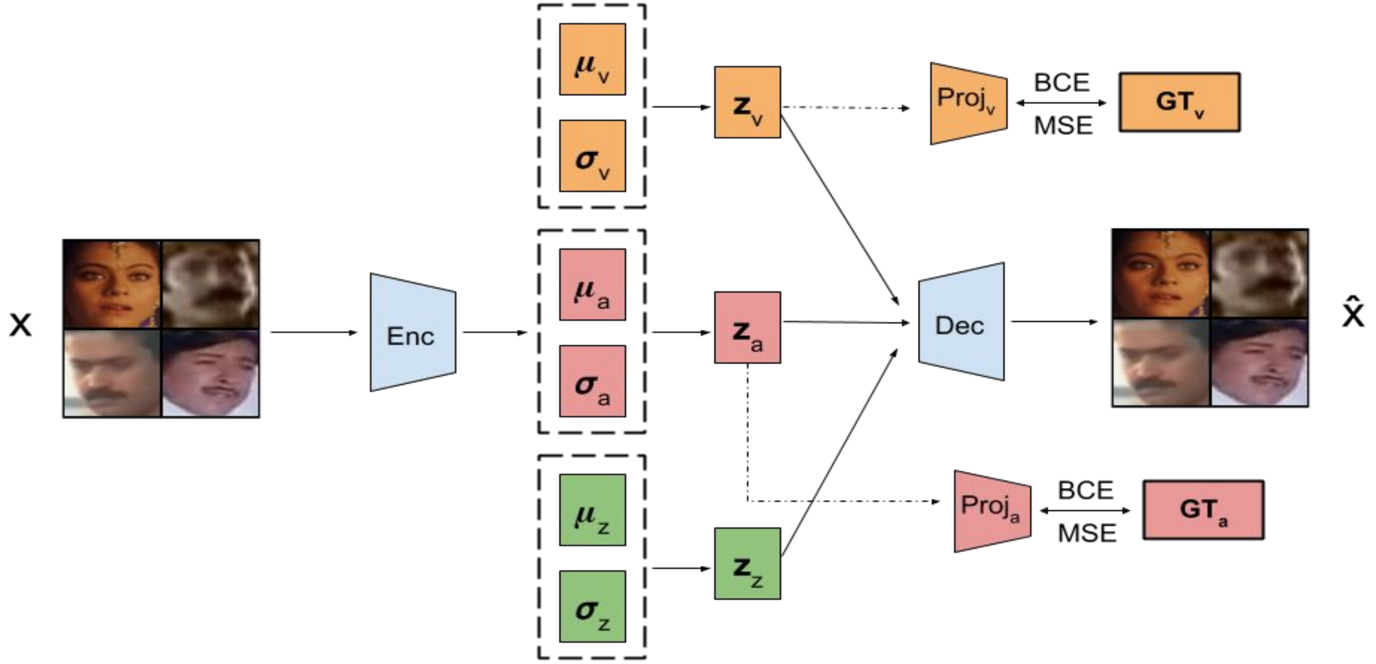


Transferred discrete VA annotations (IMFDB)

Methodology

Vanilla and Regularized VAE Training:

1. Both architectures retained the same encoder decoder structure containing 4 conv/deconv layers, batch normalization and ReLU activation
2. Models were trained with (regularized VAE) and without (vanilla VAE) regularization loss, and alignments with VA space was compared
3. For the discrete label case, BCE regularization loss was applied, while for the continuous label case, MSE regularization loss was applied
4. The models were trained for around 100 epochs on vanilla/regularized IMFDB, and around 30 epochs on vanilla/regularized AFEW
5. The performance was measured through 7 evaluation tasks



Loss = MSE(x, $\hat{x}$) + KL(z_v) + KL(z_a) + KL(z_z) + BCE/MSE($Proj_v$(z_v), GT_v) + BCE/MSE($Proj_a$(z_a), GT_a)

Vanilla VAE Regularized VAE

Network Architecture

Methodology

Evaluation Tasks

1. **Task 1:** Visualize reconstruction on the datasets for vanilla, regularized VAE.
2. **Task 2:** Measure alignment of VA space and converged VAE latent space using MSE/MAE scores between (z_v, ground truth V) and (z_a, ground truth A) for vanilla, regularized VAE.
3. **Task 3:** Compare predictive power of latent codes (z_v, z_a) for vanilla, regularized VAE, for the task of discrete emotion (seven emotions) label prediction. The task is summarized as (z_v/z_z) -> GT
4. **Task 4:** Assess regressive power of the latent codes z_v and z_a for vanilla, regularized VAE, for the task of valence regression and arousal regression. The task is summarized as z_v -> V and z_a -> A.

Methodology

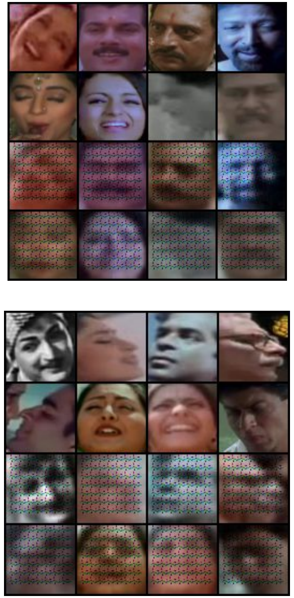
Evaluation Tasks

5. **Task 5:** Further compare the inverse regressive power of the latent codes z_a and z_v, for the vanilla and regularized VAE, for the task of inverse VA regression. The task is summarised as z_v -> A and z_a -> V.
6. **Task 6:** Visualize the latent spaces for the vanilla and regularized VAE, by plotting t-SNEs of z_v and z_a, colour-coded by their discrete emotion labels
7. **Task 7:** Visually inspect the latent spaces formed by the vanilla and the regularized VAE and model it as a circumplex space and comparing it with the GT VA annotations.

Results

Task 1: Reconstructions

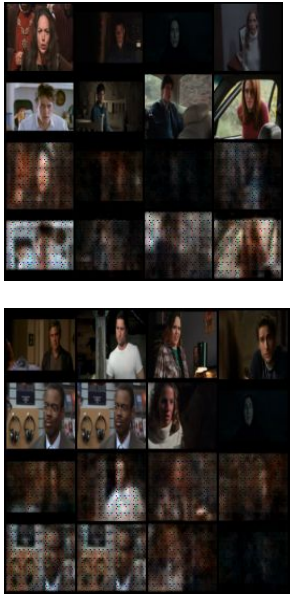
IMFDB Vanilla VAE



IMFDB Continuous Regularized VAE

IMFDB Discrete Regularized VAE

AFEW Vanilla VAE



AFEW Discrete Regularized VAE

Results

Task 2: Alignment

IMFDB Conti	V only MSE	V only MAE	A only MSE	A only MAE	Combined MSE	Combined MAE
Vanilla VAE	7.306	1.161	5.961	1.058	13.267	2.219
Reg VAE	0.573	0.582	0.257	0.367	0.83	0.95

IMFDB/AFEW Disc	Combined CE
Reg VAE	8.25/2.54

Results

Task 3: Emotion Predictive Power

Random performance: 0.142

Latent Code	DT	SVM	FC
Vanilla z_v	0.207	0.367	0.314
Vanilla z_a	0.184	0.323	0.294
Reg z_v	0.221	0.337	0.342
Reg z_a	0.209	0.307	0.326

IMFDB Continuous

Latent Code	DT	SVM	FC
Vanilla z_v	0.207	0.367	0.314
Vanilla z_a	0.184	0.323	0.294
Reg z_v	0.266	0.297	0.247
Reg z_a	0.264	0.279	0.226

IMFDB Discrete

Results

Task 4a: Regressive Power Valence

IMFDB Cont	Ridge EV	Ridge R2	Lasso EV	Lasso R2	SVR EV	SVR R2	MLP EV	MLP R2
Vanilla z_v	0.005	0.004	0	-0.001	-0.05	-0.06	-0.12	-0.24
Reg z_v	0.037	0.036	0	-0.002	0.02	0.02	0.036	0.026

IMFDB/AFEW Disc	DT Acc	SVM Acc	MLP Acc
Vanilla z_v	0.071/0.276	0.106/0.276	0.084/0.276
Reg z_v	0.084/0.282	0.12/0.282	0.108/0.283

Results

Task 4b: Regressive Power Arousal

IMFDB Cont	Ridge EV	Ridge R2	Lasso EV	Lasso R2	SVR EV	SVR R2	MLP EV	MLP R2
Vanilla z_a	-0.015	-0.018	0	-0.002	-0.069	-0.075	-0.25	-0.27
Reg z_a	0.009	0.009	0	-0.0001	0.017	-0.035	-0.062	-0.112

IMFDB/AFEW Disc	DT Acc	SVM Acc	MLP Acc
Vanilla z_a	0.118/0.153	0.211/0.153	0.171/0.134
Reg z_a	0.124/0.181	0.212/0.181	0.168/0.181

Results

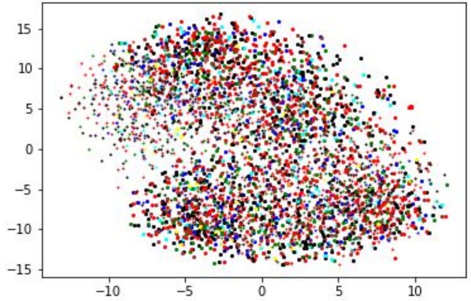
Task 5: Inverse Regressive Power

IMFDB Cont	Ridge EV	Ridge R2	Lasso EV	Lasso R2	SVR EV	SVR R2	MLP EV	MLP R2
Vanilla z_v	-0.022	-0.024	0	-0.001	-0.07	-0.09	-0.22	-0.25
Vanilla z_a	-0.01	-0.012	0	-0.001	-0.12	-0.13	-0.19	-0.22
Reg z_v	-0.015	-0.015	0	-0.001	-0.072	-0.08	-0.045	-0.045
Reg z_a	-0.015	-0.003	0	-0.002	-0.057	-0.057	-0.003	-0.057

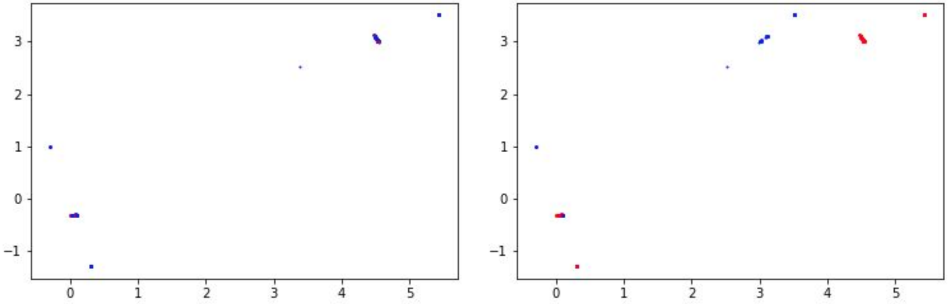
IMFDB/AFEW Disc	DT Acc	SVM Acc	MLP Acc
Vanilla z_v	0.154/0.153	0.209/0.153	0.186/0.154
Vanilla z_a	0.074/0.276	0.089/0.276	0.109/0.276
Reg z_v	0.118/0.183	0.212/0.183	0.164/0.183
Reg z_a	0.0748/0.282	0.109/0.282	0.103/0.282

Results

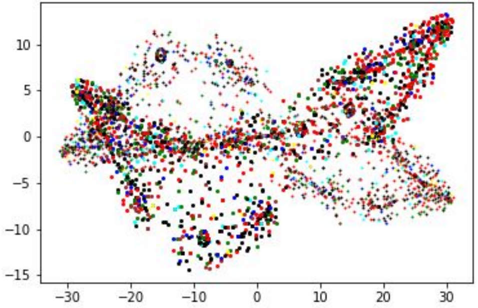
Task 6: TSNEs



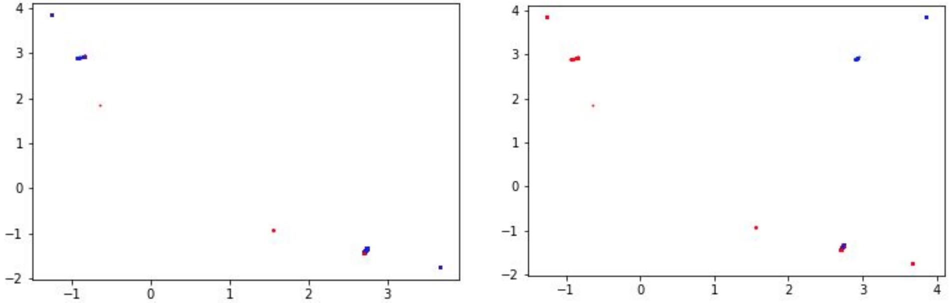
IMFDB Continuous Vanilla VAE



AFEW Discrete Vanilla VAE



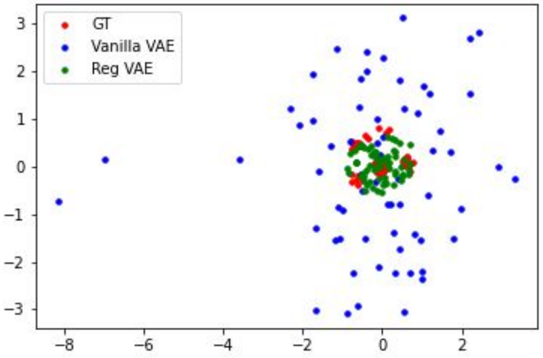
IMFDB Continuous Regularized VAE



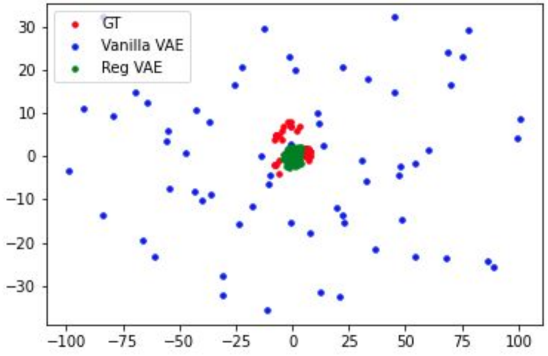
AFEW Discrete Regularized VAE

Results

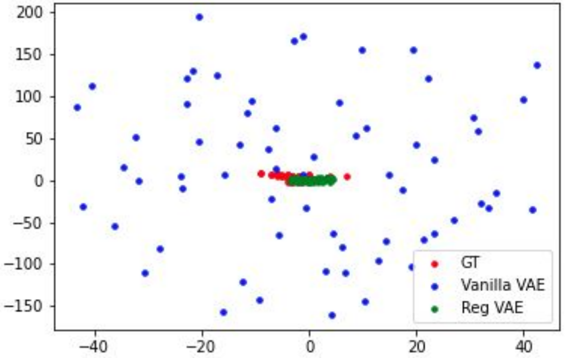
Task 7: Circumplex Representation



IMFDB Continuous



IMFDB Discrete



AFEW Discrete

Discussion

- **Task 1:** The reconstructions show that the quality of the reconstructed faces may be slightly compromised in case of regularized VAE. This can be attributed to Shannon rate-distortion theory.
- **Task 2:** Comparing divergence values for IMFDB continuous labels, we find that from a distance-metric perspective, the performance on regularized VAE increases significantly
- **Task 3:** The predictive capability of V or A alone in determining the emotion class label is low since there is less correlation between V-label and A-label
- **Task 4:** The performance of Regularized VAE is better in predicting the V and A values from their corresponding latent chunks.

Discussion

- **Task 5:** The inverse regressive performance should be poor assuming V and A are uncorrelated. This is consistently observed in the results.
- **Task 6:** In the continuous case, although not colourwise separated, the TSNE for regularized VAE is more structured as compared to vanilla VAE.
- **Task 7:** The circumplex representations clearly depict how the regularized VAE is more superior in approximation of the true VA values as compared to the vanilla VAE.

Conclusion

- The latent embeddings z_v and z_a of the regularized VAE have more predictive power with respect to the V and A annotations respectively, further highlighting the alignment of the latent space with respect to the circumplex model.
- The continuous regularized VAE lead to more structured and aligned latent spaces, corresponding to the circumplex model.

Reproducibility

- Code available here
- Saved models, results and figures are available here

References

1. S. Latif, R. Rana, J. Qadir, and J. Epps. (2017). “Variational autoencoders for learning latent representations of speech emotion.” [Online]. Available: https://arxiv.org/abs/1712.08708
2. Mina Marmpena, Angelica Lim, Torbjørn S. Dahl, and Nikolas Hemion. 2019. Generating robotic emotional body language with variational autoencoders. In Proceedings of the 8th International Conference on Affective Computing and Intelligent Interaction (ACII 2019) (ACII). IEEE.
3. Michael Suguitan, Randy Gomez, and Guy Hoffman. 2020. MoveAE: Modifying Affective Robot Movements Using Classifying Variational Autoencoders. In Proceedings of the 15th ACM/IEEE International Conference on Human-robot Interaction (HRI '20). IEEE Press.
4. A. Lindt, P. Barros, H. Siqueira and S. Wermter, "Facial Expression Editing with Continuous Emotion Labels," *2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019)*, Lille, France, 2019, pp. 1-8, doi: 10.1109/FG.2019.8756558.
5. D. Kollias, S. Cheng, E. Ververas, I. Kotsia, and S. Zafeiriou. (2018). ‘‘Generating faces for affect analysis.’’ [Online]. Available: https://arxiv.org/abs/1811.05027