Data-driven Dimensional Expression Generation via Encapsulated Variational Auto-Encoders

Wenjun Bai¹ · Changqin Quan² · Zhi-Wei Luo²

Received: 21 February 2019 / Accepted: 26 November 2021
© The Author(s) 2022

Abstract

Concerning facial expression generation, relying on the sheer volume of training data, recent advances on generative models allow high-quality generation of facial expressions free of the laborious facial expression annotating procedure. However, these generative processes have limited relevance to the psychological conceptualised dimensional plane, i.e., the Arousal-Valence two-dimensional plane, resulting in the generation of psychological uninterpretable facial expressions. For this, in this research, we seek to present a novel generative model, targeting learning the psychological compatible (low-dimensional) representations of facial expressions to permit the generation of facial expressions along the psychological conceptualised Arousal-Valence dimensions. To generate Arousal-Valence compatible facial expressions, we resort to a novel form of the data-driven generative model, i.e., the encapsulated variational auto-encoders (EVAE), which is consisted of two connected variational auto-encoders. Two harnessed variational auto-encoders in our EVAE model are concatenated with a tuneable continuous hyper-parameter, which bounds the learning of EVAE. Since this tuneable hyper-parameter, along with the linearly sampled inputs, largely determine the process of generating facial expressions, we hypothesise the correspondence between continuous scales on the hyper-parameter and sampled inputs, and the psychological conceptualised Arousal-Valence dimensions. For empirical validations, two public released facial expression datasets, e.g., the Frey faces and FERF-DB datasets, were employed here to evaluate the dimensional generative performance of our proposed EVAE. Across two datasets, the generated facial expressions along our two hypothesised continuous scales were observed in consistent with the psychological conceptualised Arousal-Valence dimensions. Applied our proposed EVAE model to the Frey faces and FERF-DB facial expression datasets, we demonstrate the feasibility of generating facial expressions along with the conceptualised Arousal-Valence dimensions. In conclusion, to generate facial expressions along the psychological conceptualised Arousal-Valence dimensions, we propose a novel type of generative model, i.e., encapsulated variational auto-encoders (EVAE), allowing the generation process to be disentangled into two tuneable continuous factors. Validated in two publicly available facial expression datasets, we demonstrate the association between these factors and Arousal-Valence dimensions in facial expression generation, deriving the data-driven Arousal-Valence plane in affective computing. Despite its embryonic stage, our research may shed light on the prospect of continuous, dimensional affective computing.

Keywords Affective computing · Generative model · Facial expression generation · Dimensional model of emotion

Introduction

Dimensional Expression Generation

Facial expression is one of the primary channels in human beings to express emotions and sentiments. How to reflect the complexity of human facial expressions is one of the central enquiries in affective computing. Lengthy researches on affective generation paid overwhelmed attention on the discrete, non-continuous, non-dimensional modelling of human affect [1, 2]. These learned simplified affective

✉ Wenjun Bai
wjbai@atr.jp

¹ Department of Computational Brain Imaging, Advanced Telecommunication Research Institute International, 2-2-2 Seika, Hikaridai, Soraku District, Kyoto, Japan

² Department of Computational Science, Kobe University, 1-1 Rokkodai-Cho, Nada-ku, Kobe, Japan

Fig. 1 Theoretical A-V dimensions of facial expressions. This theoretical A-V two-dimensional plane is the reproduction of the work in [6, 7]. The horizontal axis represents the arousal dimension, whereas the vertical axis stands for the valence dimension. The majority of human emotions can be represented on this plane. E.g., in comparison to the feeling of excitement, the feeling of happiness is relatively low in the arousal but high in the valence dimension

representations quickly reached their bottlenecks in reflecting the richness of sentiments [3]. I.e., emotions and sentiments in the wild are not restricted in certain fixed taxonomy, e.g., the basic six emotions [4]. In contrast, human facial expressions should be represented and generated as granularity at continuous dimensions. A recent neurological finding also suggests that human emotion may represent in a distributed dimensional manner [5].

Regarding the dimensional representation of facial expressions, the classic psychological conceptualised Arousal - Valence (A-V) dimensions hold great promise in defining most facial expressions along these two dimensions [6, 7]. As shown in Fig. 1, the arousal dimension measures the global feeling of dynamism or lethargy during the expressing of a certain emotion, shown as the vertical axis in Fig. 1, whereas the valence dimension offers a rough measure of the polarity of the feeling. However, in affective computing, facial expression generation along these A-V dimensions in a data-driven manner remains opaque, i.e., whether these psychological conceptualised A-V dimensions can be mathematically modelled by two continuous factors in dimensional facial expression generation is unknown.

Overview of Our Approach

To tackle the challenge of dimensional facial expression generation, among various deep learning models [8], we rely on a recent advent in generative modelling:

the variational auto-encoder (VAE) [9] to generate facial expressions in a data-driven manner. Unfortunately, it is imprudent to employ a VAE directly in generating our targeted dimensional facial expressions as it hinges on the following issue. Under the conventional framework of a VAE, the encoded latent representations of facial expressions are largely entangled. This poses the difficulty to tease our aimed two continuous factors for the arousal and valence dimension apart from the entangled latent representations. Hence, a novel form of VAE that allows the encoding of disentangled latent representations and deriving two continuous scales that correspond to the arousal and valence dimensions, is on-demand.

To this end, we propose a novel form of variational auto-encoder: the encapsulated variational auto-encoders (EVAE) that is tailored specifically for learning data-driven dimensional affective representations of facial expressions. These learned dimensional representations can then be used in generating artificial facial expressions along with the theoretical conceptualised A-V dimensions. The crux of our proposed EVAE is a tuneable hyper-parameter in permitting the encoding of flexible latent representations. These flexible, anatomical-aware latent representations can be further boiled down to two continuous factors: the included hyper-parameter and the sampled inputs for generating latent variables. Mapping these two factors onto two continuous scales allows us to establish a Cartesian coordinate system to model the theoretical A-V dimensions.

The rest article is formatted as follows. Before delineating our method, we firstly present a succinct review of previous attempts in dimensional expression generation, followed by a detailed description of our proposed encapsulated variational auto-encoders (EVAE). The succeeding section is devoted to giving a detailed description of the usage of our proposed EVAE in generating expressions along the modelled A-V dimensions. To validate our approach, two empirical experiments were conducted to demonstrate the feasibility of generating data-driven A-V dimensional expressions on two publicly available datasets, e.g., the Frey-face and FERF-DB datasets. We conclude this article by highlighting some bearing limitations and several future research avenues.

Related Works

Previous efforts in learning dimensional affective representations can be acknowledged as the degrading of the original sentimental classification problem to a regression problem in reporting continuous ratings(values) in replacement of the discrete labels [10, 11]. This trend calls for the release of large-scale dimensional annotated affect datasets, e.g., Aff-Wild [12], SEMINE [13], and AFFECTNET [14]. Relying on these datasets, several previous researches were able to revise the original discrete based classification models to regression ones that outputted the continuous arousal and valence values [15, 16]. Recent advancements along this direction is to mix the discrete categorisation with the dimensional annotations via the convex combination [17] or the multi-task learning strategy [18]. Other researches resort on multi-modal feature fusion [19, 20], rigorous defined facial features(or facial landmarks) [21, 22], and recovered 3D facial features [23] to learn dimensional affective representations.

However, the above-mentioned attempts can hardly be perceived as the distal solution in learning dimensional affective representations. Their relied annotations are expensive to collect, subjective, prone to errors, and varied across datasets. Hence, in our research, we are up against using these ambiguous annotations in developing a fully unsupervised approach to learn dimensional representations.

Using supervised self-organisation maps and a variant of neural networks, i.e., extreme learning machines, Bugnon et al [24] also aims to link the theoretical A-V space with the mathematical defined latent/graphic space. Similar to our work, their attempt is also focused on yielding a low dimensional(2D) A-V plane. Interestingly, our hyper-parameter modulated inner space between two-component VAEs can be interpreted as a latent space model, which is similar to self-organisation map. Notwithstanding, our work diverges from theirs on the generative property. The learned

latent space from EVAE with a tuneable hyper-parameter in our work allows the easy generation of the affect in accordance with A-V dimensions.

Narrowing our focus on prevailing methodological modifications on VAE, a plethora of variants that seek to learn the interpretable representation of data can be seen on the improved generalisation, e.g., beta-VAE [25] and Multi-entity VAE [26], and the usage of discretised latent space, e.g., vq-VAE [27]. Interestingly, despite the similar learning objective, i.e., to diversify the generation process of VAE, our proposed EVAE considerably differs from mentioned alternatives for its flexible independence assumption between two sets of latent variables z_s and z_b . This flexible independence constraint cannot be realised by the tuning of a hyper-parameter β in beta-VAE, the heuristic posterior parameter selection in Multi-entity VAE, or resorting to a less-accurate discretisation approach in vq-VAE.

Encapsulated Variational Auto-encoders

EVAE: Overview

Different to the conventional formation of a VAE (Fig. 2(I)), in our proposed EVAE (Fig. 2(II)), the employment of two latent variables implies a novel structured encoder-decoder architecture. I.e., for a given dataset, e.g., $x \in X$, two separate probabilistic encoders, e.g., $q_{\phi_b}(z_b|x)$ and $q_{\phi_s}(z_s|x)$ and decoders, e.g., $p_{\theta_b}(x|z_b)$ and $p_{\theta_s}(x|z_s)$ compiles up to the coinage of two component VAEs, i.e., denoting as VAE_b (the base VAE) and VAE_s (the scaffolding VAE), with latent variables z_b, z_s and two sets of to-be-optimised parameters, e.g., $\{\theta_b, \phi_b\}$ and $\{\theta_s, \phi_s\}$.

Importantly, these two latent variables z_b, z_s allow the coinage of the joint encoder and encoder of our proposed EVAE: $q_{\phi_b, \phi_s}((z_b, z_s)|x)$ and $p_{\theta_b, \theta_s}(x|(z_b, z_s))$. However, it is essential to note here that **we do not assume the conditional independency of two latent variables**. I.e., the simplified assumption of factorised joint encoder, e.g., $q_{\phi_b, \phi_s}((z_b, z_s)|x) = q_{\phi_b}(z_b|x) \cdot q_{\phi_s}(z_s|x)$ or factorised joint decoders does not hold in this research. Instead, we introduce a hyper-parameter λ to tune the relation of two latent variables in deriving its corresponded specific analytic expressions of the joint encoder and decoder.

EVAE: Encoders and Decoders

Prior to derive the analytic form of our targeted joint encoder and decoder in EVAE, we chiefly introduce the used parameterisations of the component encoders, e.g., $q_{\phi_b}(z_b|x)$ and $q_{\phi_s}(z_s|x)$ and decoders, e.g., $p_{\theta_b}(\tilde{x}_b|z_b)$ and $p_{\theta_s}(\tilde{x}_s|z_s)$.

In this case, we parameterise the base encoder, e.g., $q_{\phi_b}(z_b|x)$ under a simplified multivariate Gaussian

Fig. 2 Graphic model renderings of the conventional variational auto-encoder (VAE) and our proposed encapsulated variational auto-encoders (EVAE). (I) a conventional VAE. (II) our proposed EVAE. Solid arrows denote probabilistic decoders, whereas dash arrows represent encoders

distribution as: $\log q_{\phi_b}(z_b|x) = \log \mathcal{N}(z_b; \mu_b, \sigma^2)$, where the to-be-optimised variational parameters are ϕ_b , which can be used to produce the mean (μ_b) and s.d. (σ^2) of the approximating distribution.

Differing to the simplified parameterisation that is used in base encoder, we let a more complex full rank Gaussian distribution to be the form of our scaffolding encoder as follows, i.e., $\log q_{\phi_s}(z_s|x) = \log \mathcal{N}(z_s; \mu_s, L)$.

Where the parameterised variational parameter ϕ_s is the concatenation of the mean vector and the decomposed covariance matrix, i.e., $\{\mu_s, L\}$. Here, we use the numerical stable Cholesky decomposition to decompose the correlation matrix, e.g., Σ , into two lower triangular matrices, i.e., $\Sigma = LL^T$, to speed up the variational inference. The LKJ prior captures the covariance matrix. The LKJ prior is well suited for modelling the covariance in a multivariate normal distribution, where $x \sim \mathcal{N}(\mu, \Sigma^{-1})$. It provides a prior on correlation matrix of the parameter, i.e., $C = \text{Corr}(x_i, x_j)$, which fuses all the standard deviation of every components. We then use the numerical stable Cholesky decomposition to decompose the correlation matrix, i.e., square matrix, Σ , to be $\Sigma = LL^T$, where L is a lower triangular matrix. The LKJ distribution has the density function, e.g., $f(C|\eta) \propto |C|^{\eta-1}$.

We further let two decoders, i.e., $p_{\theta_b}(\tilde{x}_b|z_b)$ and $p_{\theta_s}(\tilde{x}_s|z_s)$, to take the similar forms of multivariate Gaussian that are used in forming the preceding encoders. I.e., $\log p_{\theta_b}(\tilde{x}_b|z_b) = \log \mathcal{N}(x; \mu_b, \sigma^2)$ and $\log p_{\theta_s}(\tilde{x}_s|z_s) = \log \mathcal{N}(x; \mu_s, L)$, where $\tilde{x}_b$ and $\tilde{x}_s$ are reconstructed inputs from the base and scaffolding VAEs respectively.

EVAE: the Importance of λ

Armed with the parameterisations on two-component encoders and decoders, it is positioned to derive the analytic expressions of our joint encoder and decoder. An optimal analytic expression of joint encoder or decoder should satisfy two following requirements. (1) The derived expression should be flexible enough to imply full factorised, full equality and the mixed relation between two-component VAEs. (2) The derived expression needs to be fully differentiable to allow fast and accurate approximate marginal inference on the variable x .

To this end, we consider the analytic expression of our targeted joint encoder that allows smooth interpolation of two-component encoders in the following Eq. 1:

$$q((z_b, z_s)|x; \phi_b, \phi_s, \lambda) \propto q_{\phi_b}(z_b|x) \cdot q_{\phi_s}(z_s|x) \cdot \mathcal{R}_e(z_b^l, z_s^l), \quad (1)$$

where the introduced discrepancy function, e.g., $\mathcal{R}_e(z_b, z_s)$ can be further expressed as:

$$\mathcal{R}_e(z_b, z_s) = \lambda \cdot \frac{1}{L} \sum_{l=1}^L \exp\left\{-\frac{1}{2\lambda^2} \|z_b^l - z_s^l\|_2^2\right\}. \quad (2)$$

As the samples are more easy to work with, we further Monte-Carlo sample these encoded representations from two encoders, i.e., $z_b^l \sim q_{\phi_b}(z_b|x)$ and $z_s^l \sim q_{\phi_s}(z_s|x)$, where l stands for the number of samples. In practise, a single sample suffices.

Interestingly, as we pay close attention to this discrepancy function in Eq. 2, the included positive factor λ offers us a tool to directly exert control over encoded latent representations.

In specific, as $\lambda \rightarrow 0$, it implies the equality constraint to enforce the two encoded latent representations from two encoders to coincide with each other. This implied equality constraint degrades our proposed EVAE to two independent VAEs. Numerically, this imposed equality constraint is equivalent with inserting a δ function to promote the exactness of two encoded latent representation as: $q_{\phi_b}(z_b|x)(\frac{1}{L} \sum_{l=1}^L \delta(z_b^l - z_s^l))$.

In opposition to the preceding case, when $\lambda \rightarrow +\infty$, two encoded latent representations are learned to diverge to the final independent relation. In this case, it assumes factorised joint encoder, i.e., $q_{\phi_b}(z_b|x) \cdot q_{\phi_s}(z_s|x)$.

Similar to the prior noted factorisation in the case of joint encoder, the analytic expression of the joint decoder, e.g., $p_{\theta_b, \theta_s}(\tilde{x}|(z_b, z_s))$ can be derived into:

$$p_{\theta_b, \theta_s}(\tilde{x}|(z_b, z_s); \lambda) \propto p_{\theta_b}(\tilde{x}_b|z_b) \cdot p_{\theta_s}(\tilde{x}_s|z_s) \cdot \mathcal{R}_d(\tilde{x}_b, \tilde{x}_s), \quad (3)$$

where the discrepancy function for the joint decoder, e.g., $\mathcal{R}_d(\tilde{x}_b, \tilde{x}_s)$ is defined as

$$\mathcal{R}_d(\tilde{x}_b, \tilde{x}_s) = \lambda \cdot \frac{1}{L} \sum_{l=1}^L \exp\{||\tilde{x}_b - \tilde{x}_s||_2^2\}. \quad (4)$$

Here, $\tilde{x}_b, \tilde{x}_s$ represents the reconstructed data instances from VAE_b and VAE_s respectively, i.e., $\tilde{x}_b \sim p_{\theta_b}(x|z_b)$ and $\tilde{x}_s \sim p_{\theta_s}(x|z_s)$. Notice here, we incorporate the same positive defined hyper-parameter λ to scale the differences of two reconstructed data instances.

EVAE: the Learning Objective

We complete our EVAE model by defining a simple factorised joint prior on two latent variables, i.e.,

$$p_{(\theta_b, \theta_s)}(z_b, z_s) = p_{\theta_b}(z_b) \cdot p_{\theta_s}(z_s). \quad (5)$$

As the learning objective of a conventional VAE is expressed as:

$$\mathcal{L}_{VAE}(\theta, \phi, x) = -\mathbb{D}_{KL}\{q_{\phi}(z|x)||p_{\theta}(z)\} + \frac{1}{L} \sum_{l=1}^L p_{\theta}(x|z), \quad (6)$$

where the first term denotes the regularisation penalty from the variational encoder, and the second term expresses the decoder induced reconstruction loss between the generation and the original input. Following the similar derivation, and the defined analytic expressions of joint encoder (cf. Eq. 1), the joint decoder (cf. Eq. 3), and the joint prior (cf. Eq. 5), the learning objective of the EVAE is defined as:

$$\mathcal{L}_{EVAE} = \mathbb{E}_{z_b, z_s \sim q(z_b, z_s|x)} \left[\underbrace{-\mathcal{D}_{KL}\{q((z_b, z_s)|x)||p(z_b, z_s)\}}_{\text{regularisation penalty from the joint encoder}} + \underbrace{\log p(\tilde{x}|(z_b, z_s))}_{\text{combined reconstruction Loss from the joint decoder}} \right]. \quad (7)$$

In Eq. 7, the first term measures the K-L divergence between the conditional density $p(z_b, z_s|x)$ and the harnessed factorised prior $p(z_b, z_s)$. The second term in Eq. 7 computes the reconstruction error between inputted samples and reconstructed ones from the joint decoder.

EVAE: the α Bounded Learning

Relying on the proposed EVAE, we are now positioned to render out the generative process. The targeted generative process is unfolded as follows: two latent representations (values of z_b and z_s) are firstly encoded from the approximated probabilistic encoders, then these latent representations are fed to the joint decoders to generate expressions.

As the encoding of latent representations is uniformly dependent upon the value of z_b and z_s , it is critical to reveal the force that has impact on the shaping of z_b and z_s . For this, we primarily rewrite the objective function in Eq. 7 into:

$$\begin{aligned} \mathcal{L}_{EVAE}(\theta_b, \theta_s, \phi_b, \phi_s, x^{(i)}) &= \mathcal{L}_{vae}(\theta_b, \phi_b, x) + \mathcal{L}_{vae}(\theta_s, \phi_s, x) \\ &\quad + \left\{ \lambda \cdot \mathcal{R}_e(z_b^l, z_s^l) - \lambda \cdot \mathcal{R}_d(\tilde{x}_b, \tilde{x}_s) \right\} \\ &= \mathcal{L}_{VAE}(\theta_b, \phi_b, x) + \mathcal{L}_{VAE}(\theta_s, \phi_s, x) \\ &\quad + \left[\alpha \right] \left\{ \mathcal{R}_e(z_b^l, z_s^l) - \mathcal{R}_d(\tilde{x}_b, \tilde{x}_s) \right\}. \end{aligned} \quad (8)$$

To rewrite the learning objective this way is to group two discrepancy functions together. The remaining terms $\mathcal{L}_{vae_b}(\theta_b, \phi_b, x)$ and $\mathcal{L}_{vae_s}(\theta_s, \phi_s, x)$ match with the learning objectives of two conventional VAEs. As seen in Eq. 8, the included hyper-parameter λ plays the key role in shaping the learning objective. However, the current support of our incorporated hyper-parameter λ lives on semi-positive defined space, which needs to be transformed to the real coordinate to serve as a continuous axis. Hence, we resort to a simple continuous function, i.e., the log function, to attain the support for λ to live in the real coordinate space, e.g., $\mathfrak{R}$. The transformed hyper-parameter is α , i.e., $\alpha \leftarrow \log(\lambda)$.

Interestingly, this newly transformed hyper-parameter α not only preserves the appealing properties of the prior noted λ , i.e., the implied independent ($\alpha \rightarrow +\infty$) and equality ($\alpha \rightarrow 0$) relations but also allows us to sample the negative values of α to influence the encoding of latent representations. As shown in Fig. 3, in terms of the learning objective, the value of α exerts its direct control on bounding the learning of EVAE.

Fig. 3 The schematic rendering on the α bounded learning of EVAE. A negative value of α sets the upper bound (cf. the (A) area) in optimisation, whereas a positive value of α defines its lower bound (cf. the (B) area)

A-V Dimensions in EVAE

Valence Dimension

To model the valence dimension, we primarily fix the α factor to a constant number. With a fixed α and sufficient training sessions, the encoded representations can be easily produced via the sampled values of learned latent variables, e.g., z_b and z_s , allowing the production of low dimensional (2D) latent representations of a given human affect dataset. The sampling technique we adopt here to produce values of latent variables relies on running the inverse cumulative probability function (CDF) of the Gaussian on the linearly spaced unit square that ranges from -1 and 1.¹ As these linearly sampled values live on the real coordinate, we then map these values on a continuous scale, denoting as X_{sample} .

As we fix the α , different values on this continuous scale lead to the diversification of encoded latent representations that correspond to different kinds of human facial expressions. In this research, **we hypothesise this continuous scale is modelled as the valence dimension in our targeted data-driven A-V dimensions.**

Arousal Dimension

Unfortunately, with a fixed α , the exploitation of previous continuous scale can merely ensure the modelling of

the valence dimension. To model the arousal dimension demands further analysis on the second factor, the included hyper-parameter α .

In specific, driving α to values that are close to 0 enforces the encoded representations from two encoders (the base and scaffolding ones) to coincide with each other in producing two identical latent representations. **In this research, we hypothesise that conditioning on these encoded uniformed latent representations, the decoder in our proposed EVAE is able to produce neutral facial expressions.** I.e., the association between the real '0' on a Cartesian coordinate and the generated neutral human affect can be established.

Pushing α to two extremes, e.g., $\alpha \rightarrow +\infty$ and $\alpha \rightarrow -\infty$, allows the diversification and unification of the encoded latent representations. Since α now lives on the unconstrained real coordinate, tuning the α from negative to positive (except 0) should elevate the degree of the same facial expression while fixing the previously defined valence axis. **Hence, a continuous scale of α is hypothesised here to represent the data-driven arousal dimension.**

Data-driven A-V Two-dimensional Plane

Armed with the modelled arousal and valence dimensions, we can finally pile them up to build our targeted data-driven A-V dimensional plane. The first axis, i.e., the arousal dimension, is the continuous scale on the introduced hyper-parameter, i.e., α , in reflecting the level of dynamism in human expressions. The second axis, i.e., the valence dimension, is modelled as the sampled value (X_{sample}) in producing the latent variables in EVAE, delineating the polarity of the human affect. Plotting these two axes together is to form a Cartesian coordinate system that renders out data-driven A-V dimensions. This yielded data-driven A-V plane is the cornerstone of this research, which transforms a psychological conceptualised plane (shown in Fig. 4(a)) to a data-driven Cartesian coordinate plane (shown in Fig. 4(b)) in representing human facial expressions. Each learned human expression can now be represented as a real number on these data-driven A-V dimensions. More importantly, as both dimensions are modelled by the continuous scales, this permits the generation of facial expressions along these two dimensions.

Empirical Validations

We ran empirical experiments on two public available facial expression datasets: the Frey faces and FER-DB [28] datasets. The objective is to empirically validate whether our proposed EVAE with a tuneable hyper-parameter α is capable of learning A-V dimensional representations for expression

¹ Presumably, the range of sampling values can be any real numbers, in favour of simple plotting, we bound these sampling values between -1 and 1 for simplicity.

Fig. 4 The conversion between the theoretical and data-driven A-V two-dimensional planes. (a) The psychological conceptualised A-V dimensional plane. (b) The data-driven A-V dimensional plane. The sampled inputs (denoted as X_{sample}) is modelled as the valence dimension, whereas the continuous scale on $[\alpha]$ is modelled as the arousal dimension

generation. I.e., the valence dimension is associated with the sampled inputs in producing latent variables, whereas the arousal dimension is modelled as the included single tuneable hyper-parameter α .

All programs in experiments are written in python with libraries of tensorflow [29], pymc3 [30], and keras [31]. The partial code to reproduce experiments can be found on: http://github/leon_bai/VSE_emo.

Datasets

Frey Faces Dataset

The Frey faces dataset contains 1956 grey-scaled images of Brendan Frey's face taken from sequential frames of a video with the dimension of 20 x 28. Each image went through the input normalisation pre-processing step. This dataset is ideal to evaluate the dimensional generative performance of EVAE due to its assumed homogeneity in facial features, i.e., all emotional expressions were rendered out from a single person.

FERG-DB Dataset

Facial Expression Research Group Database (FERG-DB) [28] is an annotated facial expression dataset, which utilised MAYA 3D modelling software to create the expressions in animated characters. Differ from the previous Frey faces dataset, this dataset is created specifically for automatic facial analysis. The entire dataset contains 55767 annotated facial expression images of six created characters. The modelled facial expressions range from low valence expressions, e.g., anger, disgust, sadness and fearful expressions, to high valence ones, e.g., surprise and joyful expressions. To fit with our objective, we intentionally discard the original discrete annotations that associate with expressions. Each expression image in the FERG-DB dataset was grey-scaled and normalised.

Experimental Set-ups

For two encoders in our proposed EVAE, two different feed-forward neural networks with isotropic Gaussian priors, e.g., $p_{\theta_b}(x|z_b) = \mathcal{N}(z_b; 0, I)$ and $p_{\theta_s}(x|z_s) = \mathcal{N}(z_s; 0, I)$, were employed to warp the variational parameters of base and scaffolding encoders in EVAE, forming two inference networks. For the decoders, two complementary feedforward neural networks were used as recognition networks.

For Frey faces dataset, the network configurations on inference and recognition networks are summarised in the upper panel of Table 1. Note that, reckoning on other feasible configurations for the encoders and decoders, e.g., the number of hidden units and the type of layer-wise activation, the relative performance of models were observed as insensitive to these choices. For the cost function, we strictly follow the derived learning objective in Eq. 8. To account for the reconstruction loss in the cost function, we were in favour of the mean square error as the primary choice as it induced stability in the practical implementation.

Compared to the previous implementation on the Frey faces dataset, learning A-V dimensional representations on FERG-DB facial expressions is a difficult task as some non-essential facial features for expression perception are also included, e.g., the female or muscular of the faces, the different hairstyles. Hence, as for the employed inference and recognition networks, a different set of configurations was applied here, shown in the lower panel of Table 1. In terms of the loss function and the choice of gradient ascent algorithm, we continue to use the mean square loss and Adam [32], respectively. Different to previously fixed learning epochs, we adopt the early stopping technique here to prevent the overfitting via setting the patience to -2^2 .

² This corresponds to the termination of the training after two non-improving epochs of training. Therefore, the number of batches included in training was largely varied across training sessions here.

Table 1 Implementation details of EVAE on the Frey face and FERF-DB datasets

Dataset	Model Architecture & Training Configurations
Frey Faces	Input: 28 x 20 x 1 Base Encoder: Conv 32x3x3 (stride 2) Pooling(stride 2) Conv 64x3x3 (stride 2) FC 200 ReLU activation Scaffolding Encoder: Conv 32x3x3 (stride 2) Pooling(stride 2). Conv 32x3x3 (stride 2) FC 200. Gaussian ReLU activation Decoders: Deconv 32x3x3 (stride 2) Deconv 64x3x3 (stride 2) ReLU activation batch Size: 100 num of batch: 150 early stopping: No optimiser: Adam
FERF-DB	Input: 48 x 48 x 1 Base Encoder: Conv 32x3x3 (stride 2) Pooling(stride 2) Conv 64x2x2 (stride 2) FC 128 ReLU activation Scaffolding Encoder: Conv 32x2x2 (stride 2) Pooling(stride 2). Conv 32x2x2 (stride 2) FC 140. Gaussian Tahn activation Decoders: Deconv 32x2x2 (stride 2) Deconv 64x2x2 (stride 2) Tahn activation Batch Size: 128 num of batch: varied early stopping: Yes optimiser: Adam

Evaluation Metrics

In comparison with the easy assessment of discriminative models in classification and regression tasks, providing an objective metric to assess the performance of a generative model is a challenging task. To this end, both quantitative and qualitative evaluations have to be taken into the consideration. The quantitative evaluation is served to measure how well our proposed EVAE is trained, whereas the qualitative one is served for assessing the visual fidelity of generated expressions along A-V dimensions.

In terms of chosen metrics, for the quantitative measurement, we default the commonly used optimised lower bound (elbo), i.e., $\mathcal{L}$ [33], as our target index. Meanwhile, for the comparative purpose, we include the original VAE as the

benchmark in the quantitative evaluation. For the qualitative evaluation, the perceptual quality of generated expressions along the modelled A-V dimensions is our prior criteria in assessing the generative performance of our proposed EVAE.

Empirical Results

Frey Faces Dataset

Quantitative Evaluation

In terms of quantitative results on Frey faces dataset, as the learning of EVAE is bounded by the continuous scale on α , this bounding effect leads to varied optimised lower bound across different α , shown in Fig. 5.

As the optimised lower bound of EVAE is varied across different α , direct comparison between our proposed EVAE with the conventional VAE on this metric yields limited ratifications.

Qualitative Evaluation

Rather than the previous less grounded quantitative assessment, we gravitate towards the qualitative evaluation. Relying on the generative property of our proposed EVAE, tuning the magnitude along the modelled valence dimension from one end to the other, we wish to observe a shift in generated expression from negative to positive ones. This is achieved via the forward propagation of a linear spaced unit square (2D) to the inverse CDF function of the Gaussian to produce the latent variables. The latent variables were then fed to the learned recognition networks to generate the corresponding images.

As shown in Fig. 6, a clear shift in generated facial expressions from the disgust-like negative one (the leftmost one in Fig. 6) to the joy-like expression (the rightmost one in Fig. 6) was observed. The observed expression shift demonstrates the feasibility of generating expressions along the modelled valence dimension.

Facial expressions that are different in the modelled arousal dimension should exhibit a differentiated degree of variations to the same affective signal. As demonstrated in Fig. 7, along the modelled arousal dimension, varying degrees of facial expressions are generated from different instantiations of EVAEs. Facial expressions that were generated at two ends of the arousal dimension are rendered out diversified and unified patterns of generation, respectively. This observed α related variations in expression generation shows that via tuning the magnitude on the modelled arousal dimension, our proposed EVAE is capable of generating expressions in line with the theoretical arousal dimension.

Fig. 5 The α bounded learning of EVAE on the Frey faces dataset. The value of α , e.g., from negative (-50) to positive (50), has direct the impact on the optimised learning elbo. As the value of α moves towards the positive end, the optimised learning elbo is attenuated accordingly

FERG-DB Dataset

Quantitative Evaluation

Following the similar bounded learning effect on the Frey faces dataset (Fig. 5), the hyper-parameter α also bounded the learning of EVAE on FERG-DB dataset. Delineated in Fig. 8, the choice of α value has the direct impact on the final optimised learning elbo of our proposed EVAE on this dataset.

Qualitative Evaluation

To evaluate whether our proposed EVAE is capable of generating expressions along the modelled valence dimension, e.g., the sampled input, via the linear interpolation of the latent space and fixing α to a certain value, a gradual shift on generated facial expressions was observed in Fig. 9. In comparison with the previous implementation of EVAE on the Frey faces dataset, the observed expression shift on the FERG-DB dataset is rendered out in a still evident but less smoothed manner.

Fig. 6 Facial expression generation along with the valence dimension on the Frey faces dataset. In this figure, based on the learned valence axis, the linear interpolation of it allows the generation of facial expressions that smoothly transformed from the negative (low in the

valence dimension) to positive affects (high in the valence dimension). Note here, the learned valence representations were rescaled to $[-1, 1]$ scale from the original $[0, 1]$ scale for the clear demonstration

Fig. 7 Facial expression generation along with the modelled arousal dimension on the Frey faces dataset. The modelled arousal axis is depicted at the leftmost of this figure. We choose 5 most representative points on this continuous value, e.g., 50, 25, 0, -25, -50. For each value along the modelled arousal dimension, the corresponded Frey's facial expressions are generated. I.e., the negative value leads to more dramatic facial expressions, and a wider range of regenerated affect (high arousal) in comparison to its positive alternative. Setting the arousal index to 0, the generated affects are neutral expressions. Note here, the axis is rendered from the positive to negative ends. This is done in purpose to coincide with the theoretical A-V plane [6]

Fig. 8 The α bounded learning of EVAE on the FERF-DB dataset. It is clear that on the FERF-DB dataset, the hyper-parameter α can exert direct control over the elbo in optimisation

As shown in Fig. 10, through tuning the numerical value of α , our proposed EVAE is able to generate designated expressions that correspond to their rough position on the theoretical arousal dimensions. We inherit the previous defined operational definition of the arousal dimension. I.e., high in the arousal dimension should lead to the generation of diversified facial expressions, whereas low in the arousal dimension leads to the generation of indifferent ones. Arming with this operational definition, we observed the α induced differentiated patterns of generated expressions in terms of the presented expressions between the bottom and top panels in Fig. 10.

However, as the employed FERF-DB dataset has six different animation characters, these diversified facial features largely influence the perceptive quality of generated expressions, the impacts of certain facial features such as the hair style and identity of the character are hard to control under the current version of EVAE, leaving this to our future research pursuit.

Fig. 9 Facial expression generation along with the modelled valence dimension on the FERF-DB dataset. Through tuning the value on the modelled valence dimension, e.g., the sampled input, EVAE is capable of generating expressions from one polarity to the other

Fig. 10 Facial expression generation along with the modelled arousal dimension on the FERF-DB dataset. Similar to the layout in Fig. 7, in this figure, the modelled arousal dimension is positioned in the leftmost panel. We chose 5 most representative α values, e.g., 0, -10, -25, 10, 25 to demonstrate the α induced dimensional expression generation

Conclusion

In this pioneering research, we propose a novel form of variational auto-encoder: encapsulated variational auto-encoder (EVAE) to automatically learn psychological conceptualised Arousal-valence (A-V) dimensional representations of human facial expressions. As the encoding of latent representations from our proposed EVAE can be further boiled down to two continuous factors: the sampled input and a tuneable hyper-parameter, it allows the building of a Cartesian coordinate system along the A-V dimensions. Leveraging on the generative property of our proposed EVAE, it permits the generation of artificial facial expressions in accordance with the theoretical A-V dimensions.

Unfortunately, these learned dimensional representations of facial expressions suffer from two limitations that thwart their large-scale usages. The primary one is their reflected poor perceptual quality in terms of reproduced expressions based on these representations. The second one is the lack of objective validation metric in accessing the performance of learned data-driven A-V dimensions.

In tackling with foregoing noted issues, we highlight two main mitigation approaches for consideration. To improve the perceptual quality of reproduced expressions and lower the dataset dependency, a large-scale facial expression dataset that can be run on more complicated network structures should be considered. In order to yield a fair assessment on the performance of our learned data-driven A-V dimensions,

the validation on regenerated expressions on celebrated facial action unit system (FACS) [34] can also be an interesting research avenue.

Despite the mentioned limitations, within authors' awareness, it is the first time that dimensional representation of facial expressions, which are defined in cognitive science, can be mathematically modelled and learned from generative modelling. The demonstrated feasibility of generating facial expressions along with the conceptualised A-V dimensions sheds light on the future prospect of continuous, dimensional affective computing.

Regarding the future research directions, two streams that are in respect to both theoretical refinements and several real-life applications on the basis of our approach can be taken into the consideration. From the theoretical perspective, the future refinement is expected to greatly simplify the current model from the use of two full auto-encoders to the utilisation of two decoders, in which will largely reduce the number of training parameters to stabilise the training process. Standing on the real-life application view, our EVE method can be readily applied to a spectrum of domains including the software engineering on facial expression generation, and the psychological experiments, where the generated continuous facial expressions along A-V axes serve as the stimuli.

Funding This study is supported by National Natural Science Foundation of China under Grant No. 61472117.

Declarations

Ethical Approval This article does not contain any studies with human participants or animals performed by any of the authors.

Conflicts of Interest The authors declare that they have no conflicts of interest.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Gunes H, Schuller B, Pantic M, Cowie R. Emotion representation, analysis and synthesis in continuous space: A survey. In: 2011 IEEE International Conference on Automatic Face & Gesture Recognition (FG). IEEE; 2011. p. 827–834.
- Sariyandı E, Gunes H, Cavallaro A. Automatic analysis of facial affect: A survey of registration, representation, and recognition. *IEEE Trans Pattern Anal Mach Intell.* 2015;37(6):1113–33.
- Wöllmer M, Eyben F, Reiter S, Schuller B, Cox C, Douglas-Cowie E, Cowie R. Abandoning emotion classes-towards continuous emotion recognition with modelling of long-range dependencies. In: Proc. 9th Interspeech 2008 Incorp. 12th Australasian Int. Conf. on Speech Science and Technology SST 2008, Brisbane, Australia. 2008. p. 597–600.
- Ekman P. Are there basic emotions? 1992.
- Bush KA, Inman CS, Hamann S, Kilts CD, James GA. Distributed neural processing predictors of multi-dimensional properties of affect. *Front Hum Neurosci.* 2017;11:459.
- Russell JA. Reading emotions from and into faces: Resurrecting a dimensional-contextual perspective. 1997.
- Posner J, Russell JA, Peterson BS. The circumplex model of affect: An integrative approach to affective neuroscience, cognitive development, and psychopathology. *Dev Psychopathol.* 2005;17(3):715–734.
- Mahmud M, Kaiser MS, McGinnity TM, Hussain A. Deep learning in mining biological data. *Cogn Comput.* 2021;13(1):1–33.
- Kingma DP, Welling M. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114.* 2013.
- Gunes H, Pantic M. Dimensional emotion prediction from spontaneous head gestures for interaction with sensitive artificial listeners. In: International conference on intelligent virtual agents. Springer; 2010. p. 371–377.
- Nicolaou MA, Gunes H, Pantic M. Output-associative RVM regression for dimensional and continuous emotion prediction. *Image Vis Comput.* 2012;30(3):186–196.
- Zafeiriou S, Papaioannou A, Kotsia I, Nicolaou M, Zhao G. Facial affect “in-the-wild. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. 2016. p. 36–47.
- McKeown G, Valstar MF, Cowie R, Pantic M. The semaine corpus of emotionally coloured character interactions. In: Multimedia and Expo (ICME), 2010 IEEE International Conference on. IEEE; 2010. p. 1079–1084.
- Mollahosseini A, Hasani B, Mahoor MH. Affectnet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Trans Affect Comput.* 2017;10(1):18–31.
- Gunes H, Pantic M. Automatic, dimensional and continuous emotion recognition. *1(1):68–99.*
- Schuller B, Valster M, Eyben F, Cowie R, Pantic M. AVEC 2012: the continuous audio/visual emotion challenge. *ACM Press.* p. 449.
- Xia R, Liu Y. A multi-task learning framework for emotion recognition using 2d continuous space. *IEEE Transactions on Affective Computing.* 2017;1:3–14.
- Eyben F, Wöllmer M, Schuller B. A multitask approach to continuous five-dimensional affect sensing in natural speech. *ACM Transactions on Interactive Intelligent Systems (TiS).* 2012;2(1):1–29.
- Lee NN, Cu J, Suarez MT. A real-time, multimodal, and dimensional affect recognition system. In: PRICAI 2012: Trends in Artificial Intelligence, Lecture Notes in Computer Science. Springer, Berlin, Heidelberg. p. 241–249.
- Soleymani M, Asghari-Esfeden S, Pantic M, Fu Y. Continuous emotion detection using eeg signals and facial expressions. In: Multimedia and Expo (ICME), 2014 IEEE International Conference on. IEEE; 2014. p. 1–6.
- Dahmane M, Meunier J. Continuous emotion recognition using Gabor energy filters. In: Affective computing and intelligent interaction. Springer; 2011. p. 351–358.

22. Schacter D, Wang C, Nejat G, Benhabib B. A two-dimensional facial-affect estimation system for human–robot interaction using facial expression parameters. 27(4):259–273.
23. Chen H, Li J, Zhang F, Li Y, Wang H. 3d model-based continuous emotion recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015. p. 1836–1845.
24. Bugnon LA, Calvo RA, Milone DH. Dimensional affect recognition from HRV: an approach based on supervised SOM and ELM. p. 1.
25. Burgess CP, Higgins I, Pal A, Matthey L, Watters N, Desjardins G, Lerchner A. Understanding disentangling in β -vae. arXiv preprint [arXiv:1804.03599](https://arxiv.org/abs/1804.03599). 2018.
26. Nash C, Eslami SMA, Burgess C, Higgins I, Zoran D, Weber T, Battaglia P. The multi-entity variational autoencoder. In: NIPS Workshops. 2017.
27. van den Oord A, Vinyals O, Kavukcuoglu K. Neural discrete representation learning. In: Proceedings of the 31st International Conference on Neural Information Processing Systems. 2017. p. 6309–6318.
28. Aneja D, Colburn A, Faigin G, Shapiro L, Mones B. Modeling stylized character expressions via deep learning. In: Asian Conference on Computer Vision. Springer; 2016. p. 136–153.
29. Abadi M, Barham P, Chen J, Chen Z, Davis A, Dean J, Devin M, Ghemawat S, Irving G, Isard M, et al. Tensorflow: a system for large-scale machine learning. In: OSDI, vol. 16. 2016. p. 265–283.
30. Salvatier J, Wiecki TV, Fonnesbeck C. Probabilistic programming in python using pymc3. PeerJ Computer Science 2016;2:e55.
31. Chollet F et al. Keras: Deep learning library for theano and tensorflow. 2015;7(8). <https://keras.io/k>.
32. Kingma DP, Ba J. Adam: A method for stochastic optimization. arXiv preprint [arXiv:1412.6980](https://arxiv.org/abs/1412.6980). 2014.
33. White T. Sampling generative networks: Notes on a few effective techniques.
34. Ekman P, Rosenberg EL. What the face reveals: Basic and applied studies of spontaneous expression using the Facial Action Coding System (FACS). Oxford University Press, USA; 1997.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.