Gated Variational AutoEncoders: Incorporating Weak Supervision to Encourage Disentanglement

Matthew J. Vowels, Necati Cihan Camgoz, Richard Bowden
University of Surrey: CVSSP
Guildford, Surrey, UK
m.j.vowels@surrey.ac.uk

***Abstract*—Variational AutoEncoders (VAEs) provide a means to generate representational latent embeddings. Previous research has highlighted the benefits of achieving representations that are disentangled, particularly for downstream tasks. However, there is some debate about how to encourage disentanglement with VAEs, and evidence indicates that existing implementations do not achieve disentanglement consistently. The evaluation of how well a VAE's latent space has been disentangled is often evaluated against our subjective expectations of which attributes should be disentangled for a given problem. Therefore, by definition, we already have domain knowledge of what should be achieved and yet we use unsupervised approaches to achieve it. We propose a weakly-supervised approach that incorporates any available domain knowledge into the training process to form a Gated-VAE. The process involves partitioning the representational embedding and gating backpropagation. All partitions are utilised on the forward pass but gradients are backpropagated through different partitions according to selected image/target pairings. The approach can be used to modify existing VAE models such as beta-VAE, InfoVAE and DIP-VAE-II. Experiments demonstrate that using gated backpropagation, latent factors are represented in their intended partition. The approach is applied to images of faces for the purpose of disentangling head-pose from facial expression. Quantitative metrics show that using Gated-VAE improves average disentanglement, completeness and informativeness, as compared with un-gated implementations. Qualitative assessment of latent traversals demonstrate its disentanglement of head-pose from expression, even when only weak/noisy supervision is available.**

***Keywords*-VAE; disentanglement; representation learning; generative models.**

I. INTRODUCTION

Variational AutoEncoders (VAEs) have gained in popularity for the unsupervised generation of low-dimensional representational embeddings over high-dimensional distributions such as images [1]–[3]. It has been demonstrated [1], [4]–[6] that if representations are disentangled (such that each representational dimension uniquely and independently corresponds with a single generative factor), then better results are achieved in downstream tasks. However, how to achieve disentanglement is an ongoing area of research, and there is evidence that recent proposals do not achieve disentanglement consistently [7], [8].

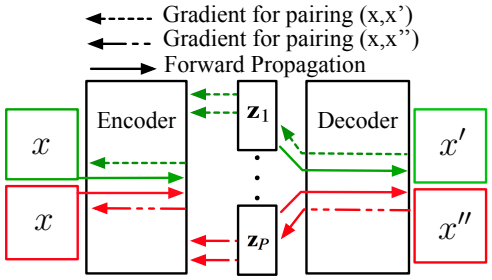

Figure 1. Gated-VAE training principle. Images are paired according to shared subsets of latent factors, where the subsets are derived from any available supervision. A forward pass is made through the whole network, but gradients are backpropagated through a specific latent partition $\boldsymbol{z}_i$ where $i = [1...P]$ according to the input/target image pairing.

We present a novel weakly-supervised approach to training VAEs which we will refer to as a Gated-VAE (see Figure 1). This involves gating the backpropagation of gradients through a partitioned latent space, where the gating is determined by the input and target image pairs. The modification can be applied to any existing VAE model. The efficacy of the Gated-VAE is demonstrated quantitatively using the disentanglement, completeness and informativeness metrics from [9] and is compared against un-gated implementations of a vae. We consider the relative importance of disentanglement versus informativeness by comparing the quantitative metrics. We also qualitatively evaluate latent traversals for images of faces where noisy/weak supervision is available.

The paper is structured as follows: In Section II we provide an overview of recent endeavours to improve disentanglement in VAEs. In Section III we cover background theory in order to provide a baseline from which to build our proposal. In Section IV we present the results from our quantitative and qualitative evaluations of Gated-VAE. Finally, a conclusion summarises the results and their significance.

II. BACKGROUND

VAEs are a type of generative model, and their performance for certain tasks is commonly compared against alternative generative models such as the Generative Adversarial

Network (GAN) [10]. GANs have been shown to provide superior generative quality, but VAEs have a number of advantages which include outlier robustness, improved training stability and interpretable, disentangled representations [7]. Disentangled representations are generally conceived to be representations in which each element relates to an independent (and usually semantically meaningful) generative factor [4], [7]. Achieving a disentangled representation is suggested to aid in downstream tasks [4], however, there is some debate [7], [8] as to whether disentanglement helps *per se*, or whether it is the *informativeness* of the latent space that primarily determines the utility of the embedding - i.e. whether the representation fundamentally captures variation in the underlying factors. Disentanglement is essential if a self-contained subspace in the full latent embedding/representation needs to be extracted or masked for downstream purposes. For example, consider the task of unsupervised facial expression representation. In the wild, most images of faces will contain some variation in head-pose (i.e. on-axis frontal images of faces would represent the exception, not the norm). In this application it may therefore be useful to derive a representation of facial expression independent of head-pose. If head-pose could be reliably disentangled from facial expression, downstream tasks that depend on facial expression could be fed with expression representations invariant to head-pose. However, VAEs do not disentangle with either predictability or consistency [8].

In recent years, various attempts have been made to encourage disentanglement with VAEs, including increasing the emphasis on reducing the distance between the posterior and the prior [4], utilisation of alternative objective functions [1], [5], [11], [12], Gaussian mixtures [13], use of alternative prior distributions [14], and cascaded vae models [7]. However, results from a review of 12,000 current implementations [8] indicate that there is almost as much influence from random initialisation as there is from hyperparameter selection and objective functions. Indeed, consistent disentanglement has recently been demonstrated to be impossible without inductive bias and subjective validation [7], [8]. In other words, the typical evaluation of the inferred latent space is subjectively compared (e.g. using reconstructions of latent traversals) against our prior expectations / domain knowledge concerning which attributes should be disentangled for a given problem. Our work incorporates any available supervision or domain knowledge into the training procedure in order to encourage disentanglement.

Beyond subjective interpretation, there is not yet a consensus on the best way to quantitatively measure disentanglement, although various proposals have been made. These include Separated Attribute Predictability [5], Mutual Information Gap [15], FactorVAE metric [11], Modularity [16], the β-vae metric [4] and the later relative of the β-vae metric [9]. Whether any of these metrics measure disentanglement as it is generally conceived is unclear [8]. The metrics proposed by [9] have been chosen for evaluation, they represent one of the most recent attempts to measure disentanglement and distinguish between disentanglement and informativeness, as well as providing an estimation of completeness (terms described in more detail in Section IV-A1). We utilise these metrics to contribute insight into the relationship between informativeness and disentanglement.

III. METHODOLOGY

This section begins with an overview of vae theory before a presentation of the currently proposed formulation.

A. Variational AutoEncoders - Background Theory

The reader is directed to [17]–[19] for a more detailed introduction to VAEs. In essence, and following the process for variational inference for a distribution of latent variables, we start by sampling from a latent distribution $\mathbf{z} \sim p(\mathbf{z})$ and generate dataset X of images $\mathbf{x} \in \mathbb{R}^N$ with observational distribution $p_\theta(\mathbf{x}|\mathbf{z})$ such that we may derive an inferred posterior for the latent distribution as $q_\phi(\mathbf{z}|\mathbf{x})$ that approximates the true conditional latent distribution $p_\theta(\mathbf{z}|\mathbf{x})$. Both $q_\phi(\mathbf{z}|\mathbf{x})$ and $p_\theta(\mathbf{x}|\mathbf{z})$ are parameterised by neural network encoder and decoder parameters ϕ and θ respectively [1], [2], [17]. The traditional approach [5], [18] involves maximisation of the Evidence Lower BOund (ELBO):

$$\max_{\theta,\phi} \mathbb{E}_{\mathbf{x}} \left[\mathcal{L}_{\mathrm{ELBO}}(x)\right] = \max_{\theta,\phi} \mathbb{E}_{\mathbf{x}} \left[\mathbb{E}_{\mathbf{z}\sim q_\phi(\mathbf{z}|\mathbf{x})} \left[\log p_\theta(\mathbf{x}|\mathbf{z})\right] - \beta \mathrm{KL}\left(q_\phi(\mathbf{z}|\mathbf{x}) \| p(\mathbf{z})\right)\right] \tag{1}$$

The first term on the RHS of Eq. 1 encourages reconstruction accuracy, and the Kullback-Liebler divergence term (weighted by parameter β [4]) acts as a regulariser, penalising approximations for $q_\phi(\mathbf{z}|\mathbf{x})$ that do not resemble the prior. The objective is therefore to maximise the marginal log-likelihood of $\mathbf{x}$ over the latent distribution $\mathbf{z}$ [4], which is assumed to be Gaussian with identity covariance $\mathbf{z} \sim \mathcal{N}(0, \mathbf{I})$. The Gaussian assumption means that Eq. 1 may be written using an analytical reduction of KL divergence [5]:

$$\max_{\theta,\phi} \mathbb{E}_{\mathbf{x}} \left[\mathcal{L}_{\mathrm{ELBO}}(x)\right] = \max_{\theta,\phi} \mathbb{E}_{\mathbf{x}} \Bigg[\mathbb{E}_{\mathbf{z}\sim q_\phi(\mathbf{z}|\mathbf{x})} \left[\log p_\theta(\mathbf{x}|\mathbf{z})\right] - \frac{\beta}{2}\Big(\sum_i \left(\left[\Sigma_\phi(\mathbf{x})\right]_{ii} - \ln\left[\mathbf{\Sigma}_\phi(\mathbf{x})\right]_{ii}\right) + \left\|\boldsymbol{\mu}_\phi(\mathbf{x})\right\|_2^2\Big)\Bigg] \tag{2}$$

In Eq. 2 the $\left[\Sigma_\phi(\mathbf{x})\right]_{ii}$ indicates the diagonal covariance, and $\boldsymbol{\mu}_\phi(\mathbf{x})$ is the mean. Both the mean and covariance are learned by the network encoder and parameterise a multivariate Gaussian that forms the inferred latent distribution $q_\phi(\mathbf{z}|\mathbf{x})$. The decoder network samples from $\mathbf{z} \sim q_\phi(\mathbf{z}|\mathbf{x})$ using the reparameterisation trick [17] such that $\mathbf{z} = \mu_\phi(\mathbf{x}) + \epsilon\sqrt{\Sigma_\phi(\mathbf{x})}$ where $\epsilon = \mathcal{N}(0, \mathbf{I})$. One interpretation of disentanglement posits that it is achieved if $q_\phi(\mathbf{z}) = \int q_\phi(\mathbf{z}|\mathbf{x})p(\mathbf{x})d\mathbf{x} = \prod_i q_i(\mathbf{z}_i)$ [5].

During VAE training, the i^{th} reconstructed example from the output of the network decoder $\hat{x}_i$ is typically compared against the i^{th} input example x_i, which is therefore also used as the target, and the marginal log likelihood is maximised by minimising a loss (e.g. binary cross-entropy) between the reconstruction and the target. Over the course of training, the VAE thereby learns decoder parameters that produce the best reconstruction, conditioned on the latent embedding of the input image. In order for the vae to infer the generative distribution $\mathbf{z}$, the ground truth latent factors must both vary between examples in the dataset, *and* take on the same value in both input and target pairs. Letting each image in X have K independent generative ground truth factors $\mathbf{v} \in \mathbb{R}^K$ i.e. $\{\mathbf{v}_k\}_{k=1}^K$ which we model using a latent generative distribution $\mathbf{z} \in \mathbb{R}^M$ such that $M \geq K$. Given that the input image x_q is usually equal to the target image x_r (as is usual for auto-encoding), the values of all ground truth generative factors in the input image are therefore the same as the values in the target image i.e. $v_{k,x_q} = v_{k,x_r}, k = 1...K$, where v_{k,x_q} is the ground truth factor k for the input image, and v_{k,x_r} is the same ground truth factor k for the target.

Readers are referred to [2] for a review of the various modifications proposed to the objective in Eq. 1 that encourage disentanglement and minimise reconstruction cost. However, as mentioned, there is some evidence to suggest that the random initialisation of the network has almost as much impact as the network architecture and objective functions have on the ability of the network to disentangle the latent space [8]. Furthermore, VAE disentanglement is often evaluated against prior subjective expectations, and therefore, by definition, we are utilising domain knowledge to evaluate the success of an unsupervised method. For instance, Higgins et al. [4] generate reconstructions of traversals/interpolations over the latent space in order to visually ascertain which of the latent space dimensions correspond e.g. with shape or rotation. Furthermore, it is often the case that some form of weak supervision is available for any given task. For example, if we have frames from a sequence, it is likely that the appearance and identity of individuals is unlikely to change between subsequent frames within that sequence. In more extreme cases, full supervision for all generative factors may be available and utilised to achieve disentanglement in an entirely prescribed way (e.g. [20]). We propose the Gated-VAE which allows domain knowledge to be employed at training in order to encourage disentanglement in a novel, weakly-supervised way.

B. Gated Variational AutoEncoders - Formulation

Often, weak supervision is available in some form (e.g. data may be clustered, or have weak labels). If any supervision is available then it should be incorporated into training in order to aid disentanglement. A Gated-VAE provides a means to incorporate available supervision into existing VAE models. The intuition behind the Gated-VAE is that input and target images can be paired according to shared factors, and that the network should learn to recognise and learn what is common between these pairs. In other words, by pairing the input image with a target image that shares specific latent factors, the network can be encouraged to disentangle these shared factors. Such pairing may be possible when weak supervision (e.g. clustering) is available. If the supervision is available then it ought to be incorporated where possible. Backpropagation of error can then be directed through specified partitions of the latent space such that different partitions are disentangled and each contains information relating to the shared factors. The approach is deemed to be weakly supervised because the input and target images need to be paired according to some prior knowledge or labels. Weak supervision is generally used to describe the scenario whereby labels are available but the labels only relate to a limited number of factors (e.g. labels may be fully supervised and describe head-pose in terms of roll, pitch and yaw, or be weakly supervised and simply indicate that two images simply share the same head-pose) [21]. Weak supervision should not be confused with semi-supervision whereby fully informative labelling is available but only for a subset of the data [22].

More concretely, we can define a subset of latent factors $\mathbf{s} \subset \mathbf{v}$ such that $\mathbf{s} \in \mathbb{R}^L$ where $L < K \leq M$. We can partition the latent space and train each partition by using input/target pairs where $x_q \neq x_r$ but where $s_{l,x_q} = s_{l,x_r}, \forall l$. The partition will learn factors s_l but not the factors in $\mathbf{v}$ that are not in $\mathbf{s}$. Designing the input/target pairs in such a way requires domain knowledge, and therefore deviates from unsupervised training to semi-supervised training.

Starting with a vae with an M-dimensional latent space, we split the latent space into P partitions (that need not be equal in size) such that $q_\phi(\mathbf{z}|\mathbf{x})$ is parameterised as follows:

$$\mathbf{z} = [\mathbf{z}_1, \mathbf{z}_2, ..., \mathbf{z}_P] = \left[\left(\mu_{\phi,1}(\mathbf{x}) + \epsilon\sqrt{\Sigma_{\phi,1}(\mathbf{x})}\right), ..., \left(\mu_{\phi,P}(\mathbf{x}) + \epsilon\sqrt{\Sigma_{\phi,P}(\mathbf{x})}\right)\right] \tag{3}$$

During forward propagation, all partitions of the latent space are used and concatenated together. Similarly, the computation of the KL divergence is also taken over the entire latent space either by concatenating the partitions and computing the loss, or by computing the loss over each partition and concatenating the loss. However, during back-propagation, the gradient is gated according to the input/target image pairing.

An example of two training iterations with P partitions is depicted in Figure 1. In the first training iteration, Image x is paired with target Image x' such that $x \neq x'$ but that x and x' have equal ground truth latent factors for a subset of all ground truth factors, and gradients are backpropagated from end-to-end but only through a specific partition of the inferred latent space. For the second training iteration

Image x is paired with target image x'' such that x and x'' also have ground truth factors that are equal for only a subset of all ground truth factors (but a different and possibly overlapping subset to the subset shared between x and x'), and gradients are backpropagated from end-to-end but through a different partition of the latent space. Note that the decoder has access to the entire latent space on the forward passes to generate the reconstruction. If the pairing of images is consistent according to the desired partitioning across the entire dataset, then the partitions will contain different factors. Even if disentanglement has not occurred *within* partitions, disentanglement will occur *between* them.

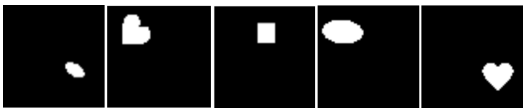

Figure 2. Samples from the dSprites [23] dataset, varying in x-position, y-position, rotation, size and shape.

In cases where reconstruction quality is desired then the training process can be split into two. First, to prioritise disentanglement using the Gated-VAE method, then by fixing the weights of the encoder and latent space, and fine-tuning the decoder by continuing its training to maximise reconstruction quality using traditional VAE training (where input image x_q is identical to target image x_r).

IV. EXPERIMENTS

The experiments demonstrate that the weakly-supervised Gated-VAE can be used to adapt existing VAE models in order to improve disentanglement. The method is first tested on synthetic data, and then on a dataset of faces.

A. Synthetic Data

We begin by demonstrating the quantitative performance improvement achieved by applying gating to three non-convolutional implementations of existing VAE models: β-VAE [4], InfoVAE [1] and DIP-VAE-II [5]. The experiments were undertaken 10 times in order to acquire averages and standard deviations for the quantitative metrics.

The dSprites [23] dataset was used for initial experiments. It comprises 737280 (64x64) images of white shapes on a black background that vary over only five generative factors: $v_0 =$ shape (square, ellipse, heart), v_1 = size (6 sizes linearly spaced), $v_2 =$ rotation (40 values over 2π), $v_3 =$ x-position (32 values) and $v_4 =$ y-position (32 values). This dataset was chosen as it is a common baseline used to test VAEs [4], [8], [24]. Samples from dSprites can be seen in Figure 2.

Input/target image pairs are chosen according to equal manifestations of generative factors. For example, in order to train the partition that is intended to learn the size of the shape in the image, a random batch of images is chosen for the input images, and the corresponding target batch is chosen such that the size of the shape in each of the target images is the same as the size of the shape in each of the paired input images. The same batching process can be applied to pair images with the same x/y position, or the same shape etc. One partition was used to represent both x and y position dimensions in order to demonstrate whether these two dimensions can be disentangled from each other within a single partition (i.e. an input image with a certain x/y position was paired with a target image with the same x/y position such that two generative factors were shared and only one partition was gated). Note that, in the case of the dSprites dataset, full supervision is available and is being used to identify the pairs. However, the labels are not provided explicitly to the network by virtue of the input/target pairing process of the Gated-VAE.

1) Evaluation - Synthetic Data: The disentanglement, completeness and informativeness metrics from [9] are used for quantitative evaluation of the Gated-VAE. These metrics build on the β-VAE metric proposed in [4] and are derived using linear Lasso and non-linear Random-Forest (RF) regressors that predict the ground truth factors using the latent embeddings of a test dataset. The regressors provide a matrix of relative importance, representing the importance of each inferred latent dimension for predicting each of the ground truth generative factors. These matrices are used to generate Hinton diagrams [25] for visualisation purposes.

According to the definitions in [4], [6], [9], *disentanglement* describes the degree to which each inferred factor in $\mathbf{z}$ independently predicts a corresponding ground truth factor. Concretely, and as defined in [9], disentanglement of inferred factor $z_i = \mu_{\phi,i}(\mathbf{x})$ is calculated as $D_i = (1 - H_K(P_i))$ where $H_K(P_i) = -\sum_{k=0}^{K-1} P_{ik} \log_K P_{ik}$ is the entropy and $P_{ij} = R_{ij} / \sum_{k=0}^{K-1} R_{ik}$ is the probability that inferred factor z_i is used by a classifier or regressor to predict ground truth factor v_j (which is a modified form to [9] to be consistent with terms in this paper). A weighting is calculated using $\rho_i = \sum_j R_{ij} / \sum ijR_{ij}$ so that the average disentanglement is weighted using $\sum_i \rho_i D_i$. These definitions mean that a disentanglement score of ≈ 0 corresponds with an inferred variable that does not contribute any predictive power.

Completeness [9] is complementary to disentanglement in that it is calculated using the same relationships for disentanglement as set out above but for each of the M inferred latent dimensions, rather than the K generative factors. A score of 1 for a generative factor v_k means that this generative factor is predicted by a single inferred latent dimension, and a score of 0 means the generative factor is predicted equally by all inferred latent dimensions.

Informativeness [9] describes whether the inferred latent representation is useful in predicting ground truth factors. In other words, it tells us whether the latent space contains useful information about the generative factors. It is quantified using the average regressor prediction error such that a lower prediction error corresponds with a higher informativeness and therefore a lower value is desirable. We

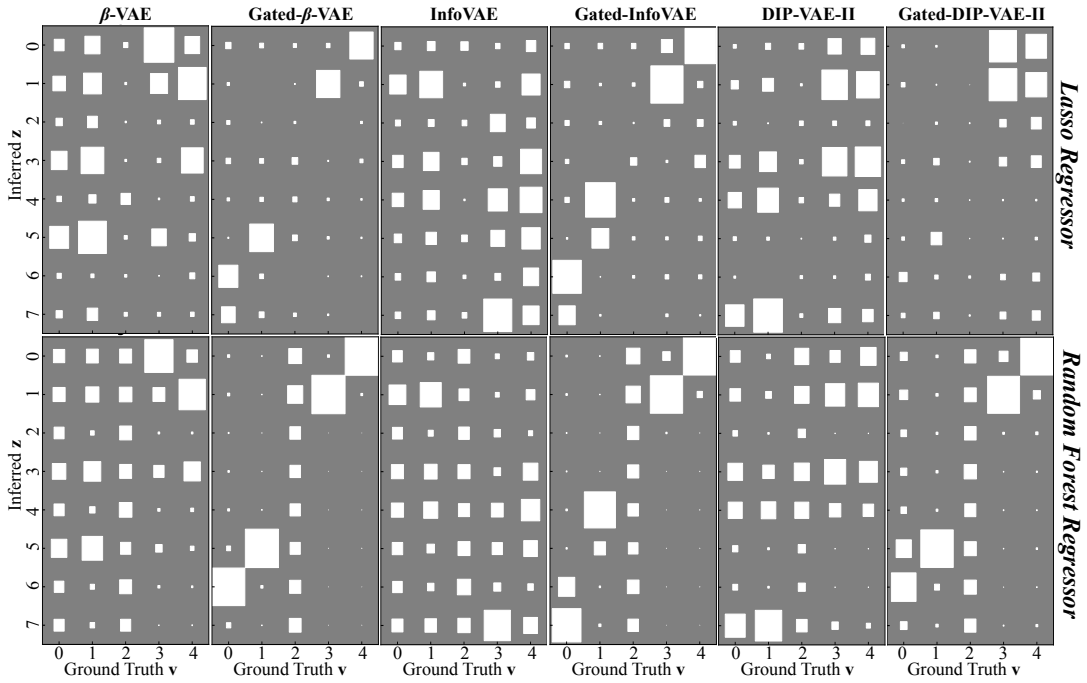

Figure 3. Example Hinton diagrams for each of the gated and un-gated models depicting the relative importance of each inferred latent dimension $\mathbf{z}$ for predicting the ground truth generative factors $\mathbf{v}$. v_0 = shape, v_1 = size, v_2 = rotation, v_3 = x-position and v_4 = y-position.

used the normalised root mean squared error.

Given these definitions for informativeness and disentanglement, informativeness becomes distinct from disentanglement, as an inferred latent representation may be highly informative without necessarily being disentangled. The characteristics of the metrics motivate the use of the RF (i.e. non-linear) regressor, because learned embeddings do not need to be linear with respect to their corresponding ground truth factors [9]. For example, any rotational factors which 'wrap' around 2π may be embedded sinusoidally. Similarly, learned embeddings capturing more than one generative factor simultaneously may be non-linearly informative, even though they are not disentangled. The results for both the Lasso regressor and RF regressors are presented in Section IV-A3. The Lasso regularisation was $\alpha = 0.02$ for all runs and the RF used 10 estimators each with a max depth of 12.

2) Models: Gating will be applied to three recently proposed variants of VAEs: β-VAE [4] (which increases the pressure on the KL-divergence loss), InfoVAE [1] (which minimises maximum mean discrepancy) and DIP-VAE-II [5] (which minimises the 2nd central moment of the latent space). Readers are referred to the original papers for a more detailed description of these models. In terms of parameter values, for β-VAE, $\beta = 4$ (as suggested by [4]), for DIP-VAE-II, $\lambda_{od} = \lambda_d = 250$, and for InfoVAE $\lambda_v = 500$ where all $\lambda\_$ parameters represent a weight on the respective component(s) of the models' objective functions.

The latent space $\mathbf{z}$ for all models had dimensionality $M = 8$ and was split into $P = 4$ partitions of equal size. $M = 8$ was chosen so that each partition could represent at most 2 generative factors, where there are 5 generative factors in total. The architecture of the basic network encoder comprises 2 fully connected layers with batch normalization and ReLU activations, and the decoder comprises 3 fully connected layers with batch normalization, ReLU activations for the first two layers and a sigmoid activation at the output. Modifications to the loss functions are made to adapt each network for β-VAE, InfoVAE and DIP-VAE-II. For all models, an Adam [26] optimiser was used with a learning rate of 0.0001, and the network was trained for 50 epochs with $N = 128$.

3) Results - Synthetic Data: Figure 3 shows example Hinton diagrams for the relative importance of each of the inferred latent dimensions $\mathbf{z}$ for predicting the ground truth factors $\mathbf{v}$, with and without our proposed gating, as well as using Lasso and RF regression. In an ideal result, the Hinton diagram would contain five distinct squares, with only one square per column and one per row. Figure 3 demonstrates consistent allocation of four out of five factors to their intended partitions in the latent space. Unfortunately, rotation (v_2) was not well encoded by *either* gated or un-gated models. The informativeness (i.e. NRMSE) of rotation for the gated and un-gated β-VAE models were 0.985 ± 0.004 and 0.964 ± 0.004 respectively. The results are similar for the other two models: RF rotation informativeness for gated and un-gated DIP-VAE-II are 0.991 ± 0.003 and 0.983 ± 0.008

Table I

AVERAGES WEIGHTED BY PREDICTOR IMPORTANCE WITH STANDARD DEVIATIONS FOR DISENTANGLEMENT, COMPLETENESS AND INFORMATIVENESS WITH AND WITHOUT GATING FOR β-VAE [4], INFOVAE [1] AND DIP-VAE-II [5] OVER 10 RUNS. NOTE: LOWER IS DESIRED FOR INFORMATIVENESS.

Regressor	Model		Disent.	Complete.	(Un)Inform.
Lasso	β-VAE	–	0.237±0.039	0.276±0.048	0.690±0.019
		Gated	**0.609**±0.136	**0.478**±0.104	**0.432**±0.087
	InfoVAE	–	0.240±0.034	0.156±0.020	0.772±0.013
		Gated	**0.647**±0.091	**0.495**±0.070	**0.481**±0.062
	DIP-VAE-II	–	0.316±0.113	0.346±0.091	0.653±0.033
		Gated	**0.487**±0.093	0.337±0.081	0.604±0.086
Random Forest	β-VAE	–	0.172±0.033	0.237±0.032	0.460±0.007
		Gated	**0.631**±0.113	**0.667**±0.093	**0.258**±0.046
	InfoVAE	–	0.113±0.030	0.142±0.026	0.483±0.004
		Gated	**0.632**±0.047	**0.646**±0.032	**0.243**±0.012
	DIP-VAE-II	–	0.197±0.097	0.346±0.095	0.456±0.008
		Gated	**0.486**±0.076	**0.527**±0.077	**0.394**±0.043

respectively; and 0.985±0.004 and 0.967±0.003 for gated and un-gated InfoVAE. The NRMSE informativeness results with the Lasso regressor are even closer to 1 for both gated and un-gated models.

The results also indicate that the x/y position factors were disentangled *within* a single latent partition. The examples shown in Figure 3 suggest that this was achieved more successfully than the un-gated equivalents. This may be because the additional supervision afforded by the Gated-VAE modification reduces the number of dimensions the VAE is disentangling at any one time.

Table I shows the results for disentanglement, completeness and informativeness for the three gated and un-gated models averaged over 10 runs using $N = 128$ for the Lasso and RF regressors. In all cases disentanglement was increased with the use of the Gated-VAE as expected. Interestingly, the informativeness of the latent space was also increased, suggesting that the Gated-VAE improves the usefulness of the learned embedding by embedding more information than the un-gated equivalents. In all cases apart from DIP-VAE-II with Lasso, gating achieved better performance in completeness than the un-gated equivalent. For DIP-VAE-II vs. Gated-DIP-VAE-II with Lasso, disentanglement was significantly improved in spite of comparable informativeness. This suggests that a space may be disentangled without being informative, or vice versa.

The consistent improvement of RF over Lasso suggests the latent representation contained information about the factors that was encoded in such a way that the non-linear RF regressor was able to fit the relationships between $\mathbf{z}$ and $\mathbf{v}$ but the Lasso regressor could not. The results demonstrate how a latent representation may be informative even if the factors are disentangled by any arbitrary degree. The improvement of the Gated-DIP-VAE-II over DIP-VAE-II was less pronounced than for the other models. This is likely to be due to the relatively poor performance of DIP-VAE-II compared with the other models, although based on previous indications of the DIP-VAE-II's performance [5], the low performance itself may be due to the particular hyperparameters used in the current experiments, and further work should involve hyperparameter optimisation of the DIP-VAE-II model.

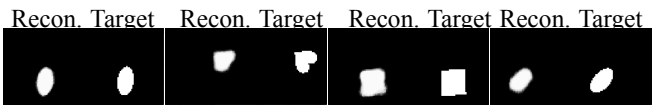

Figure 4. Reconstructions and corresponding targets for the Gated-β-VAE network with a fine-tuned decoder that demonstrate how location, size, rotation, and shape were all recovered.

Finally, Figure 4 depicts Gated-β-VAE reconstructions. The images demonstrate how, despite the poor regressor metrics, all factors (including rotation) were nonetheless encoded. In order to 'fine-tune' the network, all encoder parameters for the trained Gated-VAE were fixed, and the network decoder was then trained for 1 epoch according to the typical VAE training procedure (where the input and target images are identical).

B. Face Data

In order to demonstrate the effectiveness of Gated-VAE on more complex data, the CelebA [27] dataset is used. The CelebA dataset comprises 202,599 faces of 10,177 different individuals. The dataset was converted to greyscale to expedite training. OpenFace 2.0 [28] was used both to align the faces and also to generate labels for head-pose (pitch and yaw[1]) and facial expression (Facial Action Units - FACS), and thereby provide a source of weak supervision with which to train the Gated-VAE. The images were then clustered using K-Means, yielding 2,500 clusters for head-pose and 4,000 for facial expression. Such a clustering method is clearly not optimal if the goal is to achieve accurate labels and ideal image pairings. However, accurate supervision may rarely be available, and so this method is actually well-suited as a demonstration of how to encourage disentanglement when only noisy/weak supervision is available.

[1]No roll labels were used because OpenFace 2.0 aligns faces according to horizontal eye position and thereby removes variation in the roll dimension.

1) Evaluation - Face Data: As only weak labels (as opposed to ground truth) are available, we are unable to undertake the quantitative evaluation of performance as for synthetic data. Instead, a qualitative evaluation is performed by generating reconstructions of latent traversals. Such a method is common in the VAE literature [4], [24] and is recommended as a means of model diagnosis [29].

2) Model: Gating will be applied to a convolutional implementation of the vanilla VAE network VAE which has a weight $\beta > 1$ on the KL divergence term in the objective function). This is because increasing the β term has been shown to increase low-pass filtering characteristics (i.e. removing detail) as a side-effect of disentanglement [24]. Given also that β-VAE has been shown [8] not to disentangle latent factors in a way that is consistent (at least not with our subjective expectations) we leave disentanglement entirely to the encouragement afforded by the gating with weak-supervision. The latent space comprised two partitions of 6 (for head-pose) and 18 (for expression) dimensions.

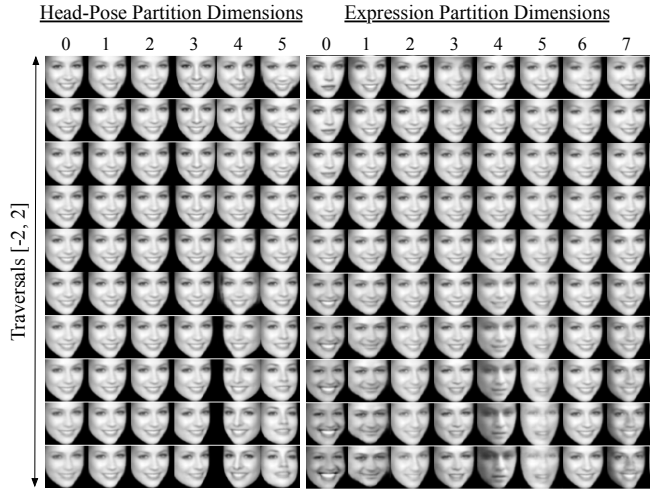

Figure 5. Reconstructions of traversals (between ± 2) of the two partitions (head-pose and expression) of the latent space for Gated-VAE. Only the 8 most active dimensions for the expression partition are shown.

3) Results - Face Data: Reconstructions of latent space traversals are shown in Figure 5. These traversals are generated by sampling and encoding a random image from the dataset, and then interpolating along each dimension step-wise between ± 2. It can be seen that head-pose has been disentangled from the rest of the generative factors, despite some degree of spill from facial expression appearing in dimension 5 of the head-pose partition. Similarly, the expression partition does not appear to contain any head-pose information, although it is interesting to note that appearance (e.g. skin colour and gender) has been encoded in this partition, despite not having been provided with supervision for this subset of factors. It can be seen that dimensions 0-2 for the head-pose partition were not used, which is as expected given that pitch and yaw can be optimally encoded with only two dimensions.

V. Conclusion

We have presented a weakly-supervised modification to the training process for VAEs which involves the partitioning of the latent space and restriction or 'gating' of gradients during backpropagation through the partitions where the gating depends on the chosen input/target image pairings. The Gated-VAE modification allows for domain knowledge to be incorporated into the training process, and can be applied to existing VAE models. The experiments compared the performance of β-VAE, DIP-VAE-II and InfoVAE with and without gating using the evaluation metrics proposed by [9], and a qualitative demonstration was presented in the form of latent traversals with the CelebA [27] dataset.

Gated versions outperformed the un-gated equivalent models in disentanglement, completeness and informativeness. The relationship between disentanglement and informativeness was illustrated, in that a disentangled latent space did not necessarily imply informativeness. The Hinton diagrams demonstrate how gating consistently allocates the relevant latent variables to a partition in the latent space thereby achieving reliable disentanglement between partitions as well as producing a more informative (and therefore more useful) latent space than the un-gated equivalent. The Hinton diagrams and individual factor results for informativeness illustrate how the rotation factor was not well encoded by either gated or un-gated models, despite being sufficiently encoded to facilitate correct reconstruction (Figure 4). This may be due to the fact that the rotation factor is lower in its salience with respect to the pixel-wise cross-entropy loss as compared with size, shape, and position. The un-gated models also had negligible informativeness for rotation. The demonstration of Gated-VAE on the CelebA dataset demonstrated that, even with noisy, weak supervision (from clustered OpenFace 2.0 output), compelling disentanglement between head-pose and facial expression was nonetheless achieved. The Gated-VAE's consistent allocation of factors to intended partitions in the latent space provides a means to mask or extract informative partitions for the purposes of downstream tasks. Further work is recommended to establish the efficacy of Gated-VAE for other downstream tasks including HCI technologies and sign language translation.

Acknowledgements

This work was funded by the SNSF Sinergia project "Scalable Multimodal Sign Language Technology for Sign Language Learning and Assessment (SMILE)" grant agreement number CRSII2_160811, the European Union's Horizon2020 research and innovation programme under grant agreement no. 762021 (Content4All) and the EPSRC project ExTOL (EP/R03298X/1).

REFERENCES

[1] S. Zhao, J. Song, and S. Ermon, "InfoVAE: Balancing learning and inference in variational autoencoders," *arXiv:1706.02262v3*, 2018.

[2] M. Tschannen, O. Bachen, and M. Lucic, "Recent advances in autoencoder-based representation learning," *arXiv:1812.05069v1*, 2018.

[3] C. K. Sonderby, T. Raiko, L. Maaloe, S. K. Sonderby, and O. Winther, "How to train deep variational autoencoders and probabilistic ladder networks," *arXiv:1602.02282v1*, 2016.

[4] I. Higgins, L. Matthey, A. Pal, C. Burgess, X. Glorot, M. Botvinick, S. Mohamed, and A. Lerchner, "Beta-VAE: Learning basic visual concepts with a constrained variational framework," *ICLR*, 2017.

[5] A. Kumar, P. Sattigeri, and A. Balakrishnan, "Variational inference of disentangled latent concepts from unlabeled observations," *arXiv:1711.00848v3*, 2018.

[6] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE Transactions on pattern analysis and machine intelligence*, 2013.

[7] B. Dai and D. Wipf, "Diagnosing and enhancing VAE models," *arXiv:1903.05789v1*, 2019.

[8] F. Locatello, S. Bauer, M. Lucic, G. Ratsch, S. Gelly, B. Scholkopf, and B. O., "Challenging common assumptions in the unsupervised learning of disentangled representations," *arXiv:1811.12359v3*, 2019.

[9] C. Eastwood and C. K. I. Williams, "A framework for the quantitative evaluation of disentangled representations," *ICLR Conference*, 2018.

[10] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," *arXiv:1406.2661*, 2014.

[11] H. Kim and A. Mnih, "Disentangling by factorising," *arXiv:1802.05983v2*, 2018.

[12] R. T. Q. Chen, X. Li, R. Grosse, and D. Duvenaud, "Isolating sources of disentanglement in VAEs," *arXiv:1802.04942v1*, 2018.

[13] N. Dilokthanakul, P. A. M. Mediano, M. Garnelo, M. C. H. Lee, H. Salimbeni, K. Arulkumaran, and M. Shanahan, "Deep unsupervised clustering with Gaussian mixture variational autoencoders," *arXiv:1611.02648v2*, 2017.

[14] J. M. Tomczak and M. Welling, "VAE with a VampPrior," *arXiv:1705:07120v5*, 2018.

[15] M. Chen, X. He, J. Yang, and H. Zhang, "3-d convolutional recurrent neural networks with attention model for speech emotion recognition," *IEEE Signal Processing Letters*, vol. 25, no. 10, pp. 1440–1444, 2018.

[16] K. Ridgeway and M. C. Mozer, "Learning deep disentangled embeddings with the f-statistic loss.," *Advances in Neural Information Processing Systems*, 2018.

[17] C. Doersch, "Tutorial on variational autoencoders," *arXiv:1606.05908v2*, 2016.

[18] D. P. Kingma and M. Welling, "Auto-encoding variational Bayes," *arXiv:1312.6114v10*, 2014.

[19] D. J. Rezende, S. Mohamed, and D. Wierstra, "Stochastic backpropagation and approximate inference in deep generative models," *arXiv:1401.4082*, 2014.

[20] T. D. Kulkarni, W. Whitney, P. Kohli, and J. B. Tenenbaum, "Deep convolutional inverse graphics network," *arXiv:1503.03167v4*, 2015.

[21] A. Szabo, Q. Hu, T. Portenier, and P. Favaro, "Challenges in disentangling independent factors of variation," *arXiv:1711.02245v1*, 2017.

[22] D. P. Kingma, D. J. Rezende, S. Mohamed, and M. Welling, "Semi-supervised learning with deep generative models," *arXiv:1406.5298*, 2014.

[23] L. Matthey, I. Higgins, D. Hassabis, and A. Lerchner, "dSprites: Disentanglement testing sprites dataset." https://github.com/deepmind/dsprites-dataset/, 2017.

[24] C. P. Burgess, I. Higgins, A. Pal, L. Matthey, N. Watters, G. Desjardins, and A. Lerchner, "Understanding disentangling in Beta-VAE," *arXiv:1804.03599v1*, 2018.

[25] G. E. Hinton and T. Shallice, "Lesioning an attractor network: Investigations of acquired dyslexia," *Psychological Review*, vol. 98, pp. 74–95, 1991.

[26] D. P. Kingma and J. L. Ba, "Adam: a method for stochastic optimization," *arXiv:1412.6980v9*, 2017.

[27] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," *Proceedings of ICCV*, 2015.

[28] T. Baltrusaitis, A. Zadeh, Y. C. Lim, and L.-P. Morency, "OpenFace 2.0: Facial behavior analysis toolkit," *13th IEEE International Conference on Automatic Face and Gesture Recognition*, 2018.

[29] L. Theis, A. van den Oord, and M. Bethge, "A note on the evaluation of generative models," *arXiv:1511.01844v3*, 2016.