

UNSUPERVISED LEARNING OF GLOBAL FACTORS IN DEEP GENERATIVE MODELS SUPPLEMENTARY MATERIAL

Anonymous authors

Paper under double-blind review

1 NETWORKS ARCHITECTURE

In this section we detail the architectures and parameters used for training the models exposed in the main paper. An extended overview is included in Table 1.

Table 1: Architecture, parameters and hyperparameters for all the models trained for the experiments presented in the paper.

Dataset	Pre-encoder	Architecture			Params	Hyperparams
		Local encoder	Global encoder	Decoder		
CelebA	h: 5 CNN layers Filters: 32, 32, 64, 64, 256 Stride: all 4 Padding: All 1 ReLU activation Batch normalization	ϕ_z : Linear layer: $256 \rightarrow 2d$ First half μ_z Second half $\text{diag}(\Sigma_z)$	ϕ_B : Linear layer: $256 + K \rightarrow 2g$ First half μ_B Second half $\text{diag}(\Sigma_B)$	θ_z : Linear layers: $g \rightarrow 256 \rightarrow 2d$ First half μ_z Second half $\text{diag}(\Sigma_z)$ θ_x : Linear layer: $d + g \rightarrow 256$ 5 transpose CNN layers Filters: 64, 64, 32, 32, 3 Stride: 1, 4, 4, 4, 4 Padding: 0, 1, 1, 1, 1 ReLU activation Sigmoid output	$d=20$ $g=50$ $K=20$ $\sigma_x=0.2$	
MNIST	h: Linear layer: $28 * 28 \rightarrow 256$ ReLU activation	ϕ_z : Linear layer: $256 \rightarrow 2d$ First half μ_z Second half $\text{diag}(\Sigma_z)$ ϕ_d : Linear layers: $d \rightarrow 256 \rightarrow K$ Tanh activation Softmax output	ϕ_B : Linear layer: $256 + K \rightarrow 2g$ First half μ_B Second half $\text{diag}(\Sigma_B)$	θ_z : Linear layers: $g \rightarrow 256 \rightarrow 2d$ First half μ_z Second half $\text{diag}(\Sigma_z)$ θ_x : Linear layers: $d + g \rightarrow 256 \rightarrow 28 * 28$ ReLU activation Sigmoid output	$d=10$ $g=20$ $K=10$ $\sigma_x=0.2$	
CelebA + 3D FACES		Same than for CelebA			$d=40$ $g=40$ $K=40$ $\sigma_x=0.2$	
3D Cars-3D Chairs		Same than for CelebA			$d=20$ $g=20$ $K=20$ $\sigma_x=0.2$	
3D Cars-Cars		Same than for CelebA			$d=20$ $g=50$ $K=20$ $\sigma_x=0.2$	

2 EXTENDED EXPERIMENTS

EXTENDED RESULTS FOR SECTION 4.1: UNSUPERVISED LEARNING OF GLOBAL FACTORS

With the aim at evaluating whether a fraction of the clusters inferred by UG-VAE encode visually interpretable global/local features, in Figure 1 we include the results for CelebA for $K = 20$ clusters. We observe that a considerable proportion of the clusters captures disentangled generative factors. Moreover, considering the heterogeneity and variety in the generative factors of celebA faces (up to 40 different attributes), increasing the number of clusters might lead to capture more representative faces, and thus, generative global factors modulated by β . In Figure 1, we appreciate that, apart from skin color, beard or image contrast, other generative factors controlled by the global variable are hair style (remarkable for components 9, 16, 17 or 18), sex (components 4 and 14), or background color (components 4, 16 and 17).

In order to visually remark the advantage of capturing global correlations among samples of UG-VAE wrt the cited related models, we include in Figure 2 an interpolation in the latent space of β -VAE, following the approach of experiment 4.1 in the paper. We explore the latent space from $\mathbf{z} = [-1, -1, \dots, -1]$ to $\mathbf{z} = [1, 1, \dots, 1]$, given that the prior is an isotropic Gaussian. As the reader may appreciate, only one row is included as β -VAE does not have global space. In this case, moving diagonally through the latent space start from a blond woman and ends in a brunette woman with the same angle face. Thus, the local space is in charge of encoding both content and style aspects. Although in β -VAE, authors analyze the disentanglement in each dimension of the latent space, we do not study whether each dimension of \mathbf{z} represents an interpretable generative factor in UG-VAE or not, as it is out of the scope for this work. The novelty lies on the fact that, apart from the local disentanglement, our model adds an extra point of interpretability through the disentanglement in the global space.

With the aim justifying the configuration for obtaining the samples exposed in Figure 3 of the paper (fixing d for the whole interpolation in \mathbf{z} and β spaces), we include in Figure 3 this interpolation process when we do not fix d . Hence, for each row, we sample d and interpolate \mathbf{z} for the selected component. The global interpolation remains equal, but as the reader might appreciate, the interpretability of which global information is controlled by β is hard to analyze by using this set up.

EXTENDED RESULTS FOR SECTION 4.2: DOMAIN ALIGNMENT

We include here the results of a interpolation in both the local space obtained when the number of components is $K = 1$, i. e., using the ML-VAE approach. As showed in Figure 4, when training ML-VAE with randomly grouped data, global space is not capable of capturing correlations between datasets, and the local space is in charge of encoding the transition from celebA to 3D FACES, which is performed within each row.

With the aim at reinforcing the robustness of UG-VAE in domain alignment, we include in Figure 6 the results of evaluating GMVAE with two clusters ($K = 2$) in a similar setup that in section 4.2. A map with the reduced latent space (using t-SNE) of GMVAE is included in Figure 5, where each point represents an encoded image. Figure 6 shows the interpolation between images of two different domains. As GMVAE does not have global variables, the interpolation applies only for the latent encodings in \mathbf{z} . Note that the interpolation is merely a gradual overlap between the two images. Namely, the model is not able to correlate the features of both images, regardless of their domain. On the other hand, with UG-VAE, by keeping fixed the global variable and interpolating in the local one, we maintain the domain but we translate the features of one image into the other. This analysis corroborates that the model finds this type of correlations in a clearly separated way.

EXTENDED RESULTS FOR SECTION 4.3: REPRESENTATION OF STRUCTURED NON-TRIVIAL DATA BATCHES

In this extension, we show another evaluation of the capacity of UG-VAE in capturing global structures. In this occasion, after training the model with randomly picked digits from MNIST, we compute the posterior of structured batches containing only even numbers, only odd numbers, numbers from Fibonacci series, and prime numbers. This grouped batches do not share strong generative

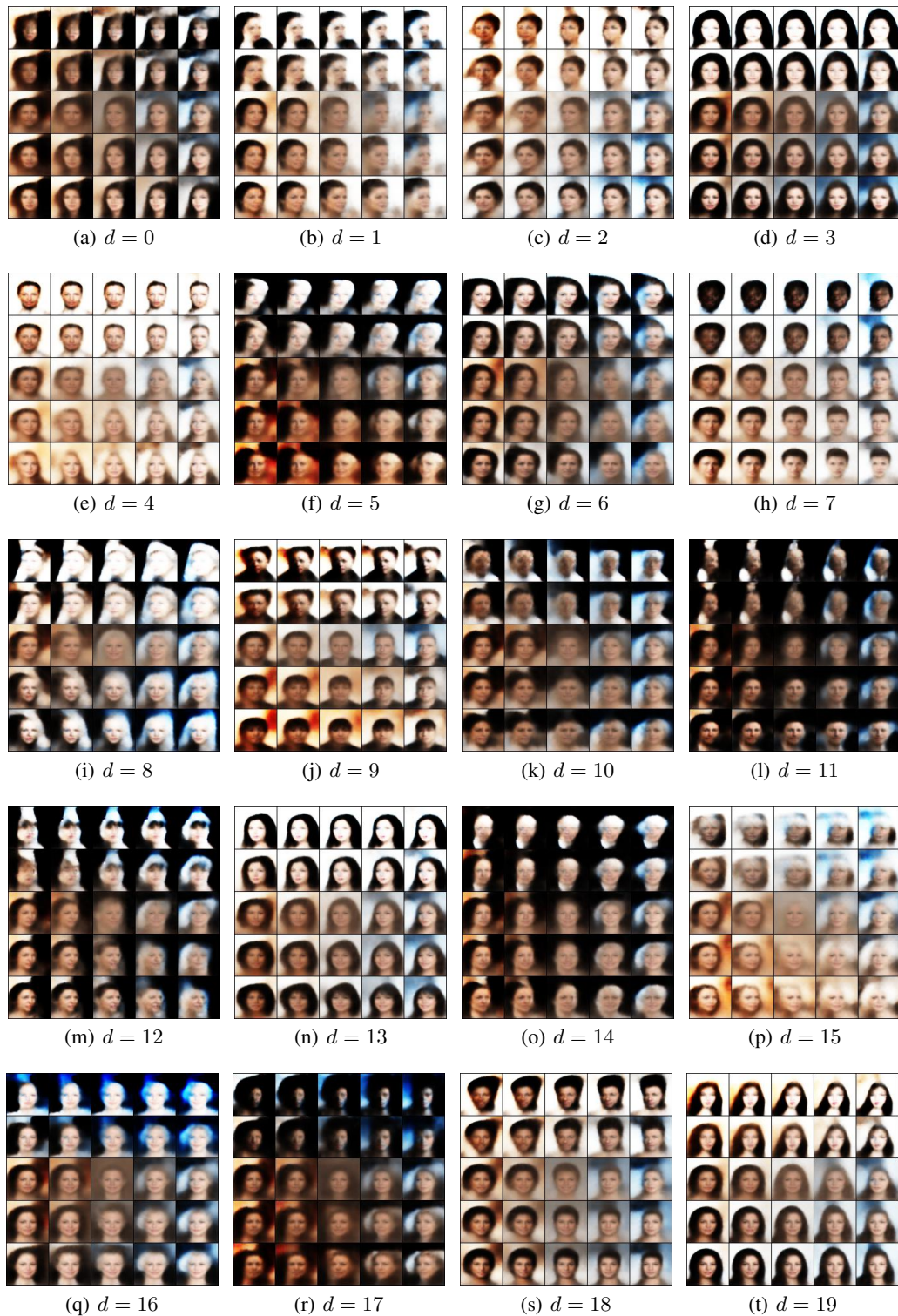


Figure 1: Sampling from UG-VAE for CelebA. We include samples from each of the $K = 20$ clusters.

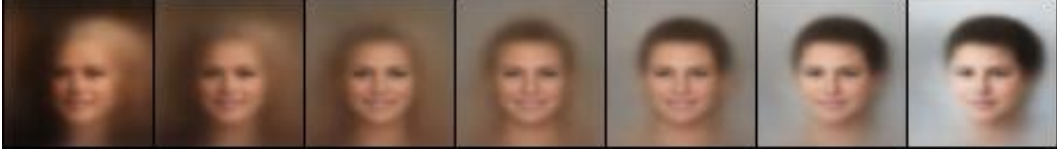


Figure 2: Interpolation in the prior latent space of β -VAE with $\beta = 10$, using the same networks architecture than in the local part of UG-VAE. Interpolation consists on 7 steps from $\mathbf{z} = [-1, -1, \dots, -1]$ to $\mathbf{z} = [1, 1, \dots, 1]$.



(a) CelebA



(b) MNIST

Figure 3: Sampling from UG-VAE for CelebA (left) and MNIST (right). We include samples for CelebA with $K = 20$ and MNIST with $K = 10$. We sample from $p(\mathbf{d})$ to obtain a cluster for each row. The information encoded in global β remains hardly interpretable by using this set up.

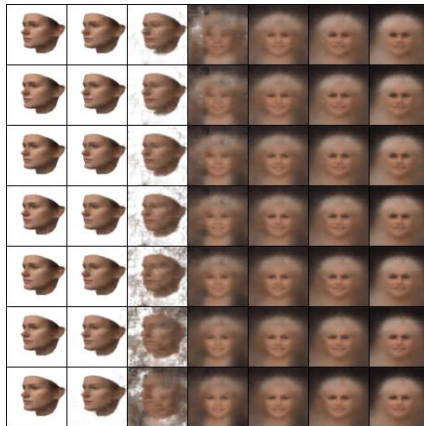


Figure 4: ML-VAE interpolation in local (columns) and global (rows) posterior spaces, fusing celebA and FACES datasets

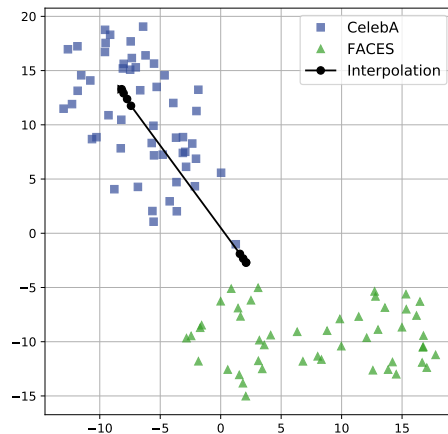


Figure 5: Interpolation map (with t-SNE) of the latent space of a GMVAE with $K = 2$ after performing domain alignment, using the same networks architecture than in the local part of UG-VAE. We interpolate between the encodings of images from CelebA and FACES dataset.



Figure 6: Interpolation in the latent space of GMVAE with $K = 2$ for performing domain alignment, using the same networks architecture than in the local part of UG-VAE. We interpolate between the encodings of images from CelebA and FACES dataset.

factors among them that influence the pixel distributions (as with CelebA groups in experiment 3.3). Namely, the only global information in this example is their frequency of appearance in each batch type. In Figure 7 we show the 2D t-SNE projection of the posterior global latent variable β distributions. We observe that UG-VAE is able to discriminate among them in the global space.

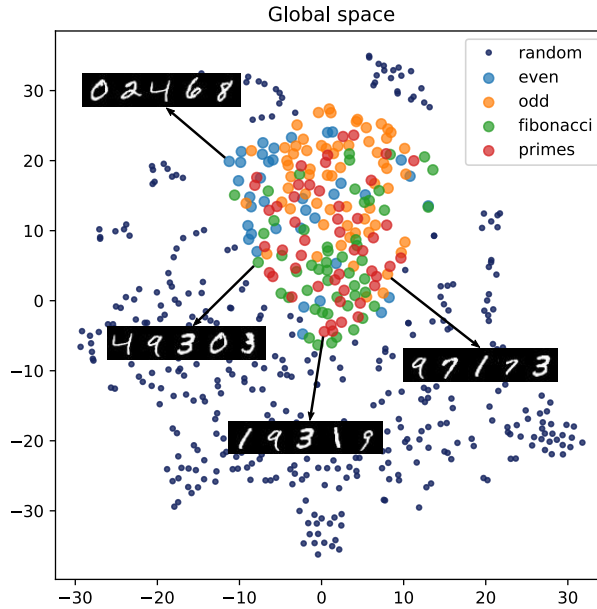


Figure 7: 2D t-SNE projection of the UG-VAE β posterior distribution of structured batches of 128 MNIST images. UG-VAE is trained with completely random batches of 128 train images.