Multilinear Latent Conditioning for Generating Unseen Attribute Combinations - Supplementary material

Markos Georgopoulos   Grigorios Chrysos   Maja Pantic   Yannis Panagakis

1. Derivation of M_{CP} in (13)

In this section we derive the final form of M_{CP} in 13. To do that, we make use of the following lemma.

Lemma 1. *It holds that*

$$(\boldsymbol{A}_1 \odot \boldsymbol{A}_2)^T \cdot (\boldsymbol{B}_1 \odot \boldsymbol{B}_2) = (\boldsymbol{A}_1^T \cdot \boldsymbol{B}_1) * (\boldsymbol{A}_2^T \cdot \boldsymbol{B}_2) \quad (1)$$

Proof. Initially, both sides of the equation have dimensions of $K \times L$, i.e., they match. The (i, j) element of the matrix product of $(\boldsymbol{A}_1^T \cdot \boldsymbol{B}_1)$ is

$$\sum_{k_1=1}^{I_1} A_{1,(k_1,i)} B_{1,(k_1,j)} \quad (2)$$

Then the (i, j) element of the right hand side (rhs) of (1) is:

$$E_{rhs} = (\sum_{k_1=1}^{I_1} A_{1,(k_1,i)} B_{1,(k_1,j)}) \cdot (\sum_{k_2=1}^{I_2} A_{2,(k_2,i)} B_{2,(k_2,j)}) = \sum_{k_1=1}^{I_1} \sum_{k_2=1}^{I_2} (A_{1,(k_1,i)} A_{2,(k_2,i)})(B_{(1,k_1,j)} B_{2,(k_2,j)}) \quad (3)$$

From the definition of Khatri-Rao, it is straightforward to obtain the (ρ, i) element with $\rho = (k_1 - 1)I_2 + k_2$, (i.e. $\rho \in [1, I_1 I_2]$) of $\boldsymbol{A}_1 \odot \boldsymbol{A}_2$ as $A_{1,(k_1,i)} A_{2,(k_2,i)}$. Similarly, the (ρ, j) element of $\boldsymbol{B}_1 \odot \boldsymbol{B}_2$ is $B_{1,(k_1,j)} B_{2,(k_2,j)}$.

The respective (i, j) element of the left hand side (lhs) of (1) is:

$$E_{lhs} = \sum_{\rho=1}^{I_1 I_2} A_{1,(k_1,i)} A_{2,(k_2,i)} B_{1,(k_1,j)} B_{2,(k_2,j)} = \sum_{k_1=1}^{I_1} \sum_{k_2=1}^{I_2} A_{1,(k_1,i)} A_{2,(k_2,i)} B_{1,(k_1,j)} B_{2,(k_2,j)} = E_{rhs} \quad (4)$$

In the last equation, we replace the sum in ρ ($\rho \in [1, I_1 I_2]$) with the equivalent sums in k_1, k_2. □

2. Additional results

2.1. Results on seen combinations

In the main paper we provide results only on unseen attribute combinations. We report results on seen combinations of MNIST and CELEB-A in Table 1. All models are able to generate samples of the seen attribute combinations. However, only the proposed models can recover the unseen combinations.

2.2. Interpolations in the latent space

To further exhibit the linear interpolation properties of our MLC-VAE model, we visualize the convex combination of all three colors. The results in Figure 1 indicate the smooth transition from one color to the other. Furthermore, it is noticeable that there is color mixing in the synthesized images, which results in the appearance of novel colors (e.g., yellow).

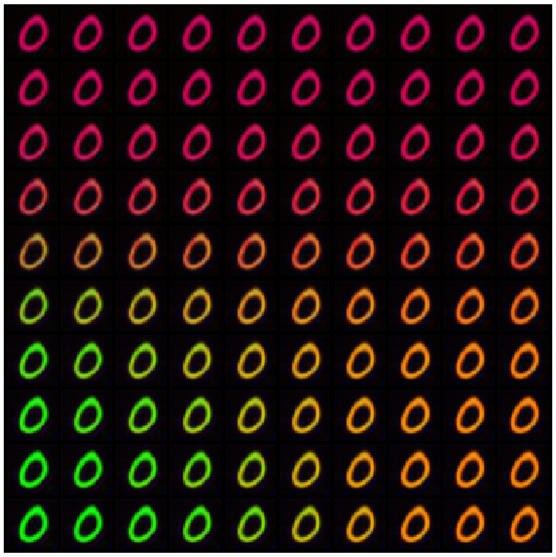

Figure 1. Three-way color interpolations in the latent space. The method is trained with three original colors (orange, magenda and green) and we interpolate between the same digit in these three colors. Notice that color mixing emerges and this results in the appearance of novel colors (e.g., yellow).

Model	**MNIST**		**CELEB-A**	
	digit	color	gender	smile
cVAE	98.17 (11.5)	**99.93** (49)	99.83 (18)	**99.97** (7.6)
VampPrior	**99.62** (2.6)	96.48 (2.3)	**100** (17.1)	**99.97** (2.8)
NCVAE-CP (Ours)	97.86 (68.1)	**99.93** (99.2)	99.7 (96.4)	99.77 (**94**)
NCVAE-Tucker (Ours)	98.24 (**95.1**)	99.89 (**100**)	99.6 (**99.4**)	99.7 (93.5)

Table 1. Accuracy (%) of the attribute classifier for each model on MNIST, FASHION, CELEB-A generation and CELEB-A transfer benchmarks on the seen (unseen) combinations.

We showcase the above quality of our framework by performing the same experiment on different colors. In particular, the model is further trained on red, green and blue digits. The results are presented in Figure 2.

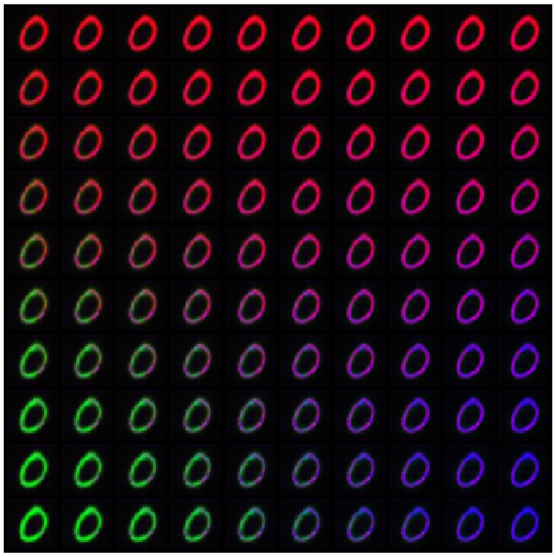

Figure 2. Three-way color interpolations in the latent space. The model is trained on red, green and blue digits. Notice that color mixing emerges and this results in the appearance of novel colors (e.g., pink).

2.3. Negative log-likelihood evaluation

We further evaluate the proposed model on CELEB-A using negative log-likelihood (NLL). Similarly to the experiments in the main paper, we train the cVAE and our model on all images except for smiling women. NLL is then reported on the images of the unseen attribute combination. In particular, NLL for cVAE was 7073.6 and for our model 6819.8.

3. Conditional VAE baseline

The label input for the conditional VAE baseline is a one-hot encoded vector. To obtain the one-hot label vector, we consider each attribute combination (y_1, y_2) a different class (cVAE-OH). A different way to encode label information from multiple attributes is to concatenate the corresponding label vectors (cVAE-concat). For example, a smiling woman in the CelebA benchmark is labelled $(y_1 = [1,0]^T, y_2 = [0,1]^T)$. Using the two aforementioned encoding schemes we get $y_{OH} = [0,1,0,0]^T$ for cVAE-OH and $y_{concat} = [1,0,0,1]^T$ for cVAE-concat. We compare the above baselines with the proposed method on the CelebA benchmark and present the results in figure 3 and table 2.

Model	**Gender**	**Smile**
cVAE-OH	18	7.6
cVAE-concat	51.2	27.3
MLC-VAE-CP	96.4	94
MLC-VAE-T	99.4	93.5

Table 2. Accuracy (%) of the smile and gender classifiers on the generated missing attribute combination.

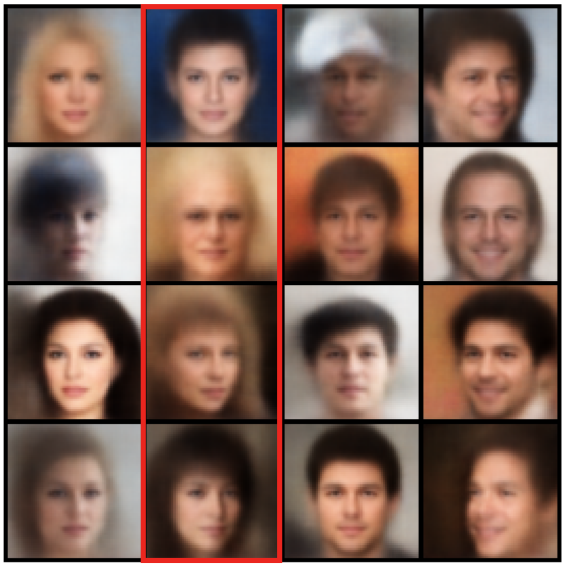

Figure 3. Synthesized samples corresponding to unseen combinations (in the red rectangle) using cVAE-concat. We notice that the generated images in the rectangle do not display smiles.

The results above highlight the inability of the baseline cVAE to synthesize unseen attribute combinations, regardless of the way we incorporate the label information.