Characteristic Capturing Variational Auto-Encoders

Hippolyte Pilchen, Victor Barbeteguy, Alexandre Cahill

Introduction to Probabilistic Graphical Models Course, **ENS Paris-Saclay**

Variational Inference for Semi-Supervision

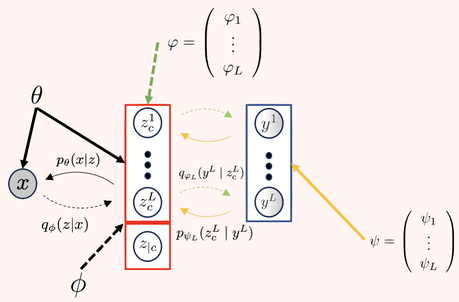
Variational Auto-Encoders proved to be a powerful framework to learn representations. This new model aims to improve the semi-supervision tasks by :

- Imposing more structure on the latent variables through a **partition** of the latent space
- Improving the precision of conditional generation through **latent disentanglement**
- Enhance the agility of **label intervention**

This is achieved by redefining the objective function and adding auxiliary tasks.

The CC-VAE Model

The CC-VAE model can be summarized by the plate below :



Alternative models

Previous work on VAE's for semi-supervision focused primarily on classification tasks. The architecures were built to improve the accuracy of the models and limit the reconstruction loss simultaneously. However this came at the cost of agility in the latent space. In order to compare the architectures, we have reconstructed the plates of different models, which we included in our report. The models at study were :

1. **M2 model :** the first semi-supervised VAE model, placing the y variable directly in the latent space. This is a major drawback as the VAE will not learn in the supervised case for that variable. Generation and intervention is also less precise as the latent variable can only be assigned a discrete number of values corresponding to the labels.
2. **DIVA** . VAE for domain-invariance learning. Similar in structure to the CC-VAE with additional classifiers assigned to the latent space. Differs from CC-VAE in the need to assign an auxillary function in the supervised case.
3. **Adapted DIVA :**. A modified DIVA model introduced in [2] for comparitive studies with CC-VAE. 2 modifications have been applied to the initial DIVA : removing the domain characteristic and functions and partitioning the latent space for labels.

Objective Functions

We conducted a comparative study of the objective functions for each model. The precise functions $\mathcal{L}_s(x, y)$ and $\mathcal{L}_u(x)$ (variational lower bounds) are mentionned in our report.

	M2	DIVA	Adapted DIVA	Our Adapted DIVA	CCVAE
$\mathcal{U}_{\text{Supervised}}$	$\sum_{i=1}^{N_s} \mathcal{L}_s(x_i, y_i) + \alpha \mathbb{E}_{p_\theta(x_i, y_i)}[-\log q_\phi(y_i \mid x_i)]$	$\sum_{i=1}^{N_s} \mathcal{L}_s(x_i, d_i, y_i) + \alpha_d \mathbb{E}_{q_\phi(z_d \mid x_i)}(log(q_{\omega_d}(d_m \mid z_d)))$ $+\alpha_y \mathbb{E}_{q_\phi(z_y \mid x_i)}(log(q_{\omega_y}(y_i \mid z_y)))$	$\sum_{i=1}^{N_s} \mathcal{L}_s(x_i, y_i)$	$\sum_{i=1}^{N_s} \mathcal{L}_s(x_i, y_i) + \alpha_y \mathbb{E}_{q_\phi(z_y \mid x_i)}(log(q_{\omega_y}(y_i \mid z_y)))$	$\sum_{i=1}^{N_s} \mathcal{L}_s(x_i, y_i)$
$\mathcal{U}_{\text{Unsupervised}}$	$\sum_{j=1}^{N_u} \mathcal{L}_u(x_j)$	$\sum_{j=1}^{N_u} \mathcal{L}_u(x_j, d_j) + \alpha_d \mathbb{E}_{q_\phi(z_d \mid x_j)}(log(q_{\omega_d}(d_j \mid z_d)))$	$\sum_{j=1}^{N_u} \mathcal{L}_u(x_j)$ $+\alpha_y \mathbb{E}_{q_\phi(z_y \mid x_j)}(log(q_{\omega_y}(y_j \mid z_y)))$	$\sum_{j=1}^{N_u} \mathcal{L}_u(x_j)$	$\sum_{j=1}^{N_u} \mathcal{L}_u(x_j)$

Table 1. Comparative study of the objective functions

Experiment 1 : Latent Embedding

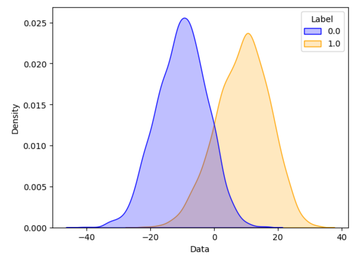


Figure 1. Latent space embedding for smiling label

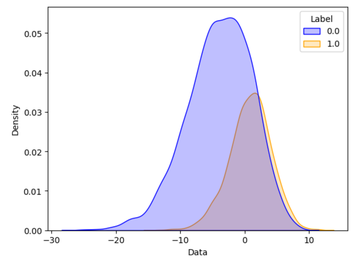


Figure 2. Latent space embedding for arched eyebrows label

Figure 3. Latent traversal for smiling label

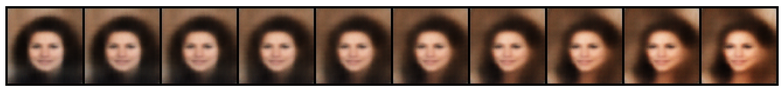
Figure 4. Latent traversal for arched eyebrows label

Figure 5. Figures 5 and Figures 6 illustrate the distributions of the embedding corresponding to "smiling" and "arched eyebrows" respectively for 10000 labeled images. We can observe that the model has learned a much clearer representation for the 'smiling' attribute where the densities for label 0 and 1 are distinguishable and clustered. This is not the case for the 'arched eyebrows' label.

Experiment 2 : Multi-class attributes

The 1D latent embeddings perform particularly poorly on labels that have been binarized but should be multi-class. For instance the model learns poorly to classify brown hair and black hair. This gave us the idea to implement a multi-class attribute for hair (containing the attributes 'Bald','Black Hair','Blond Hair','Brown Hair','Gray Hair'). However this didn't improve the performance :

- The overall **reconstruction** performance wasn't significantly higher.
- The latent space was not more **clustered** which meant that the latent space was just as entangled.
- **Conditional generation and intervention** wasn't localized to the label at study but affected the whole image.



Latent traversal in the space associated with multi-class hair label

Experiment 3 : Feature Matching

We use the *SuperGlue* model [4], a Graph Neural Network combined with Optimal Matching layers to compute keypoint correspondences between the input image and the reconstructed image.

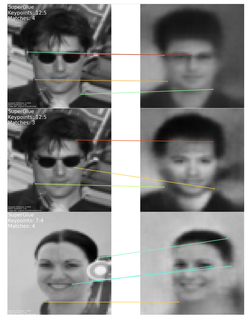


Figure 6. Examples of matching, with and without matching labels (top to bottom: with glasses, without and smiling)

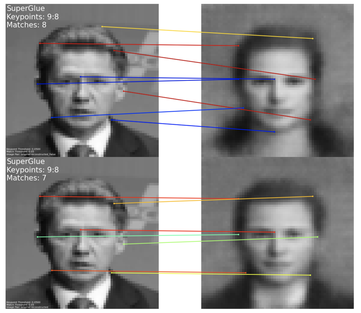


Figure 7. Feature matching modifications by intervening on *necktie* label (confidence in matching goes from blue to red)

The performance of the matching is degraded when we intervene on a single label before reconstruction. This does not only take place locally near the label but globally on the whole image. This shows that the latent space is entangled as a changing the whole part. To quantify this effect we have evaluated the mean keypoints match and the confidence of the matches for a reconstructed image with or without intervention. They both drop after intervention even though matches were all over the image and not between the labelled attributes.

Reconstruction	Mean Keypoints Match	Confidence
Reconstruction	8.88	0.76
Reconstruction + Intervention	7.74	0.62

Table 2. Feature matching results averaged on 50 images

Conclusion and limitations

By adapting the utility function and partitioning the latent space, the aim was to disentangle the latent space to make intervention and conditional generation more modular and adaptable to precise characteristics. However, our experiments showed that this was only achieved partially and came at a cost.

1. Label Selection :
   - Performs well on small hand-selected selection of labels
   - Embeddings less clustured on multi-class labels
2. High Reconstruction Cost:
   - Reconstructed images achieve poor performances, even though not initial aim of the model
3. Feature intervention exhibits latent entanglement :
   - Local intervention on a local feature applies a global transformation to the image
   - Not fully disentangled between the label latent space and the image reconstruction latent space

References

[1] Maximilian Ilse, Jakub M. Tomczak, Christos Louizos, and Max Welling. Diva: Domain invariant variational autoencoders, 2019.

[2] Tom Joy, Sebastian M. Schmon, Philip H. S. Torr, N. Siddharth, and Tom Rainforth. Capturing label characteristics in vaes, 2022.

[3] Diederik P. Kingma, Danilo J. Rezende, Shakir Mohamed, and Max Welling. Semi-supervised learning with deep generative models, 2014.

[4] Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. SuperGlue: Learning feature matching with graph neural networks. In *CVPR*, 2020.

alexandre.cahill@polytechnique.edu hippolyte.pilchen@polytechnique.edu victor.barberteguy@polytechnique.edu