Machine Learning in healthcare: Multi view Variational Auto Encoder

Filippo Nardi, Jorge Parreño Hernández, Riccardo Conforto Galli

December 2023

1 The model architecture

In a standard VAE, the objective is to maximize the ELBO, which is formulated as:

$$\text{ELBO} = E_{q_\phi(z|x)}[\log p_\theta(x|z)] - \text{KL}[q_\phi(z|x)||p(z)]$$

Here, x represents the input data, z denotes the latent variable, $p_\theta(x|z)$ is the likelihood or reconstruction probability, and $q_\phi(z|x)$ is the variational distribution, parameterized by ϕ, approximating the true posterior $p(z|x)$. The KL divergence term penalizes the discrepancy between the variational distribution and the prior distribution $p(z)$.
Now, in the context of Multi-View VAEs, suppose we have N different views of the data $(x_1, x_2, ..., x_N)$. For each view x_i, we'll have its respective encoder $q_{\phi_i}(z|x_i)$, decoder $p_{\theta_i}(x_i|z)$, and corresponding latent variable z.
The objective of a Multi-View VAE is to maximize the combined ELBO across all views:

$$\text{ELBO} = \sum_{i=1}^{N} \left(E_{q_{\phi_i}(z|x_i)}[\log p_{\theta_i}(x_i|z)] - \text{KL}\left[q_{\phi_i}(z|x_i) \parallel p(z)\right] \right)$$

Each view contributes its reconstruction term and its own KL divergence term. The variational distribution for each view aims to capture the latent space specific to that particular view. Meanwhile, the shared prior distribution $p(z)$ encourages a joint latent space where common factors across views are captured.

2 Performance

Upon evaluating the model with both MNIST and SVHN datasets as views for the encoder, the results showcase a commendable performance. The utilization of both datasets enables the model to leverage the complementary information present in these diverse views, leading to a more comprehensive and informative latent space representation. This synergy between MNIST and SVHN views contributes to improved reconstruction accuracy and better disentanglement of latent factors, reflecting the effectiveness of leveraging multiple views for learning representations.
Conversely, employing only one of the datasets – whether MNIST or SVHN – in the encoder setup yields suboptimal results. This limitation arises due to the narrower scope of information available within a single view and to reproduce a dual view. Without the complementarity offered by the diversity of views, the model struggles to capture the full spectrum of underlying data characteristics, resulting in reduced reconstruction accuracy and less effective disentanglement of latent factors.
These findings underscore the significance of leveraging multiple views in the MV-VAE framework, demonstrating that the model's performance benefits substantially from the synergy and richness provided by diverse data features.

2.1 Generating from both datasets

Figure 1: SVHN and MNIST generated from both datasets

Generating from MNIST dataset

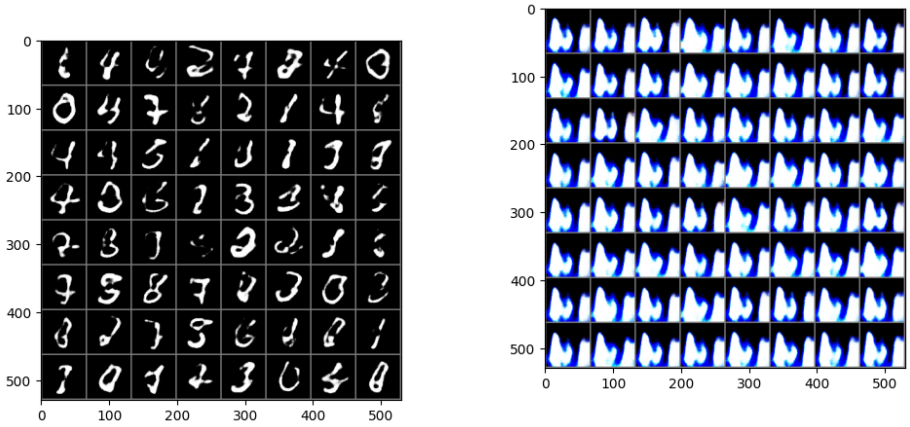

Figure 2: SVHN and MNIST generated from mnist dataset

Generating from SHVN dataset

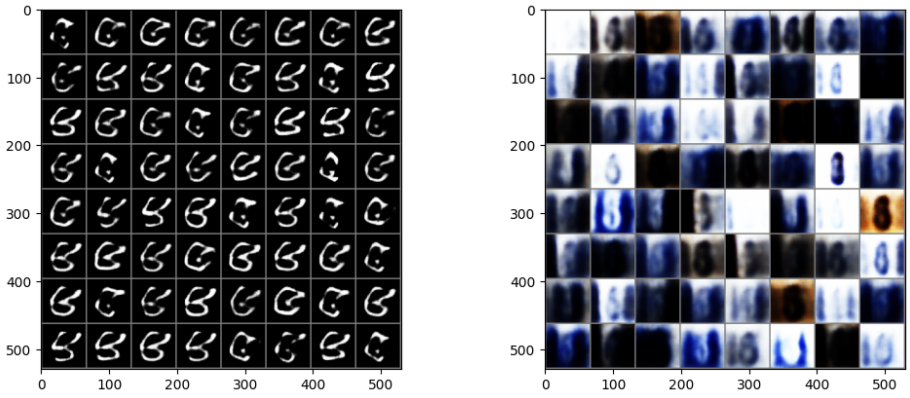

Figure 3: SVHN and MNIST generated from svhn dataset

3 Code source

The code can be found in the following github repository