Weakly Supervised Disentanglement with Triplet Network

Pedro C. C. C. C. Coutinho[1,2], Yannick Berthoumieu[2], Marc Donias[2], Sébastien Guillon[1]
TotalEnergies OneTech[1]
Université de Bordeaux, CNRS, Bordeaux INP, IMS, UMR 5218[2]

Neural networks have proven themselves useful to solve a big variety of complex tasks in many domains, such as image and neural language processing. Recently, generative models have gained considerable attention in the scientific community. Several approaches have been proposed, such as Variational Autoencoders (VAEs), Generative Adversarial Networks (GANs), and Diffusion Models. These approaches sample a latent vector $\mathbf{z}$ from a prior distribution, usually Gaussian, and transform it into new data via a trained network.

In this context, several methods have been proposed in order to attain **disentangled representations**, where each latent dimension would be linked to a single generative factor in the data generation process (Figure 1). Obtaining disentangled representations would be interesting since not only it allows us to conditionally generate new data, but also to interpret the latent space. However, training a model to find such solution is a complex task, and multiple works have proposed different methods to do so, mostly in the VAE framework.

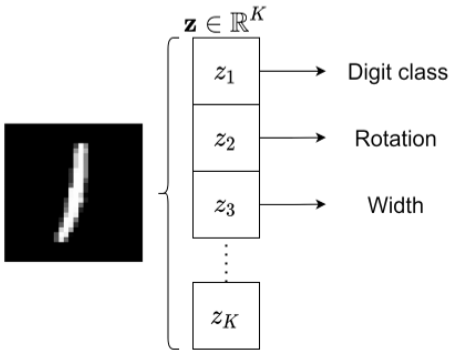

Figure 1 – Example of a disentangled representation for a MNIST sample

Hence, the objective of this work is to propose an approach that allows us to **explicitly** disentangle at least one factor of variation by introducing some weak supervision to the standard VAE in the form of a triplet loss (Figure 2). By using a triplet of images to train our network, where two of them share a common generative factor whereas the third one has a different factor, we are able to encode this specified factor into a single group of latent space dimensions.

Experiments carried out on MNIST and dSprites dataset have allowed us to demonstrate these concepts.

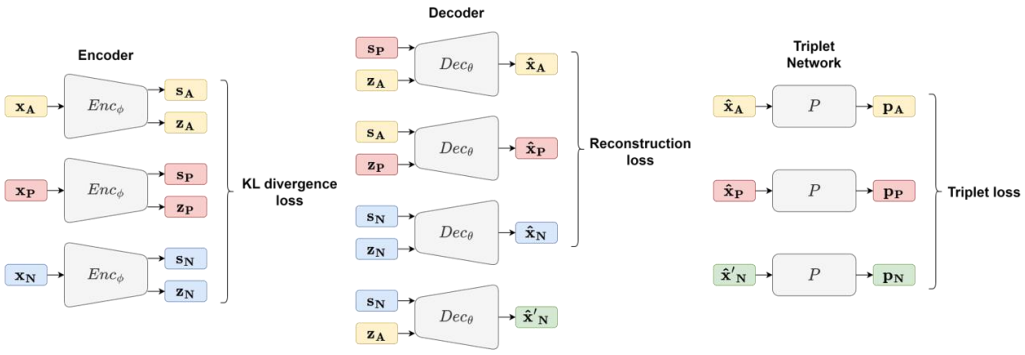

Figure 2 - Proposed method. The VAE is trained with a triplet of images, where $\boldsymbol{x_A}$ and $\boldsymbol{x_P}$ share the same generative factor, whereas $\boldsymbol{x_N}$ has a different one. The triplet network should be able to predict such factor by its ouput $\boldsymbol{p}$ such that $\boldsymbol{p_A}$ is close to $\boldsymbol{p_p}$, but far from $\boldsymbol{p_N}$