group of machine
gmum
learning research

Generative AutoEncoder

J. Tabor (S. Knop, P. Spurek, I. Podolak, M. Mazur)

Instytut Informatyki i Matematyki Komputerowej UJ

gmum.net

Generative model

group of machine
gmum
learning research

- Data are given as $x \in X$,
- Model is to learn to generate new data from the true probability distribution.

Generative model – how to do it wrong

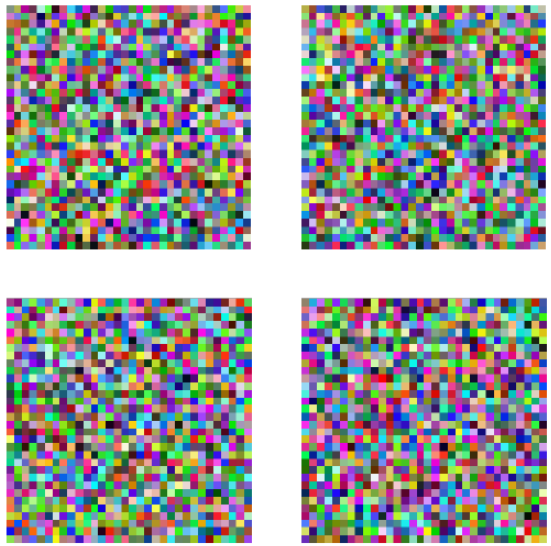
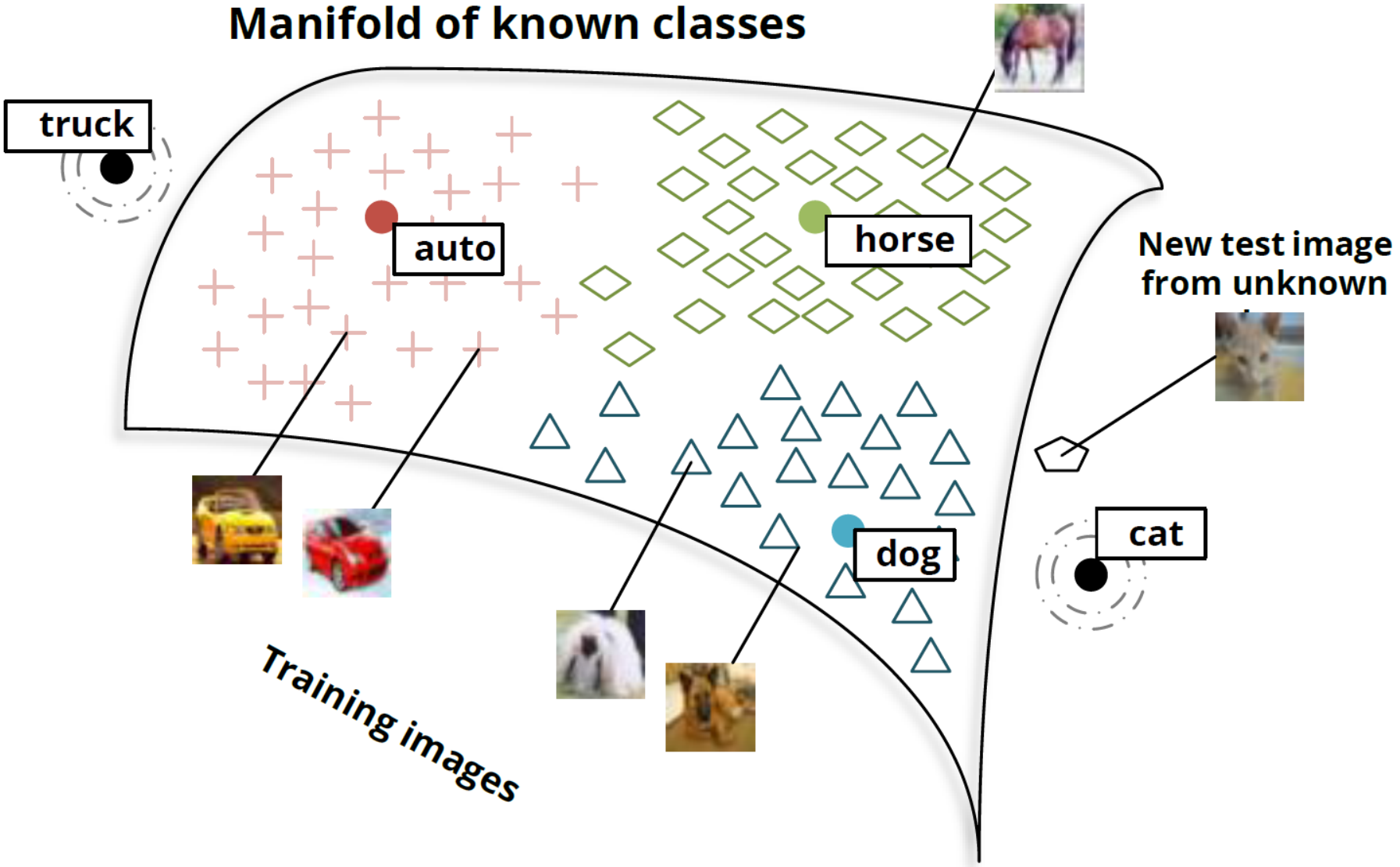


Figure: Example images drawn from a uniform distribution

- Images drawn from a uniform distribution do not correspond to anything we may meet in the real space
- Uniform distribution incorrectly models the true data
- Meaningful images take up only a very small part of the whole X space – we say that they lay on a low dimensional manifold

Generative model – how to do it right

Generative model – how to do it right



Generative model

group of machine
gmum
learning research

Generative model performs two tasks at the same time

- selects the low dimensional manifold
- computes which data are more or less frequent

Generative model: how to use - example

We can apply arithmetic (and more general) operations on data.

group of machine gmum learning research

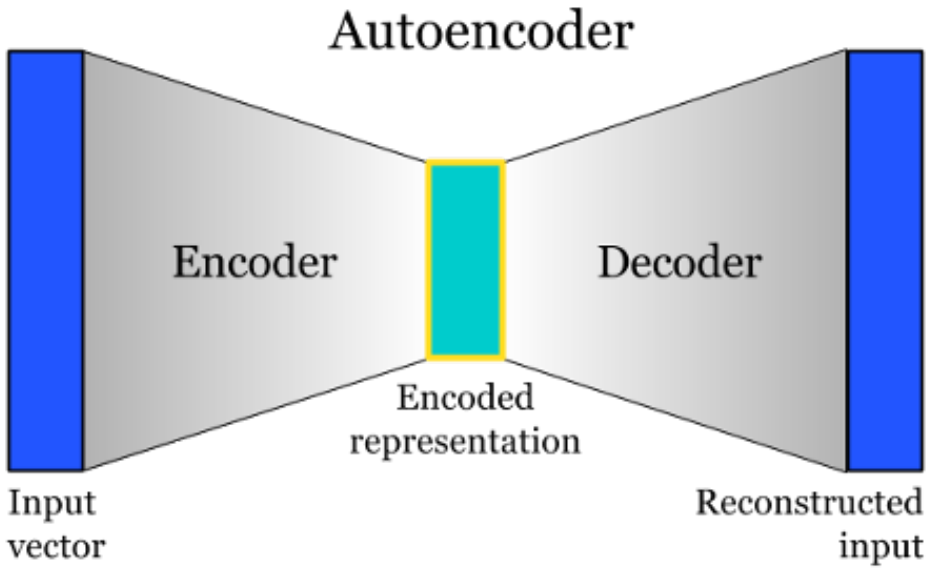
https://www.youtube.com/watch?v=-R9bJGNHltQ

Construction of the manifold - AutoEncoder

group of machine gmum learning research

Generalization of PCA, idea based on compression of dataset $X = (x_i) \subset \mathbb{R}^N$ to a linear space Z of smaller dimension D (*latent space*).

We have an encoder $\mathbb{R}^N \ni x \to \mathcal{E}x \in Z$ and decoder $Z \ni z \to \mathcal{D}z \in \mathbb{R}^N$. We want to find such encoder and decoder which minimize reconstruction error: $\text{Rec\_Error} = \sum_i \|x_i - \mathcal{D}(\mathcal{E}x_i)\|^2$.



AutoEncoder gives us the lower dimensional manifold on which the data lies (but no distribution).

Generative AutoEncoder

group of machine learning research

The aim is to ensure, that the data transported to the latent space comes from the standard normal distribution $\mathcal{N}(0, I)$.

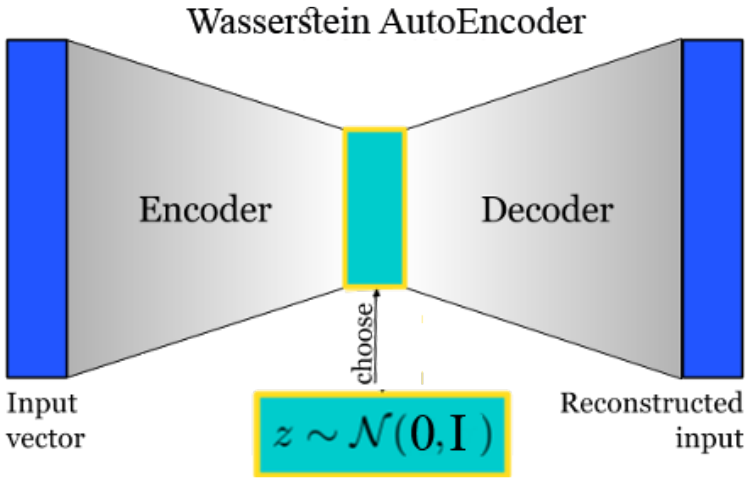


Figure: Generative AutoEncoder.

Then we can sample from our distribution by sampling from $\mathcal{N}(0, I)$ in the latent and transporting by the decoder to the input space.

BHEP normality test

group of machine
gmum
learning research

Generally considered as the best normality test is the BHEP. It measures the L^2 distance between the regularized sample (by kernel density approach) and regularized normal density [1]:

$$T_{n,\gamma} = \| N(0, I + \gamma I) - \frac{1}{n} \sum_{i=1}^{n} N(x_i, \gamma I) \|_{L_2}^2, \tag{1}$$

where γ is the smoothing parameter.

BHEP works well in small dimensions $D \leq 5$, but as show our experiments fails for large dimensions and standard sample size, since the reliable kernel density estimation in high dimensions needs extremely large samples [2, Subsection 4.5].

BHEP as metric on distributions

group of machine
gmum
learning research

One can observe that

$$T_{n,\gamma} = \| N(0, I + \gamma I) - \frac{1}{n} \sum_{i=1}^{n} N(x_i, \gamma I) \|_{L_2}^2$$

$$= \| [N(0, I)] * N(0, \gamma I) - [\tfrac{1}{n} \sum_{i=1}^{n} \delta_{x_i}] * N(0, \gamma I) \|_{L_2}^2$$

BHEP as metric on distributions

group of machine gmum learning research

One can observe that

$$T_{n,\gamma} = \|N(0, I + \gamma I) - \frac{1}{n}\sum_{i=1}^{n} N(x_i, \gamma I)\|^2_{L_2}$$

$$= \|[N(0, I)] * N(0, \gamma I) - [\frac{1}{n}\sum_{i=1}^{n} \delta_{x_i}] * N(0, \gamma I)\|^2_{L_2}$$

Consider two distributions μ, ν on $\mathbb{R}^D$. Then we can consider the BHEP test as the metric, which is defined by as the L_2-distance between regularized distributions:

$$d^2_\gamma(\mu, \nu) := \|\mu * N(0, \gamma I) - \nu * N(0, \gamma I)\|^2_{L_2}.$$

where γ is the smoothing parameter ($\gamma = \frac{1}{2\beta^2}$). As mentioned before, works well in small dimensions $D \leq 5$.

Sliced approach to comparison of distributions

group of machine
gmum
learning research

By the Cramer-Wold Theorem (also Radon Transform) we can compare two distributions by comparing all one-dimensional projections. Given a density f and $v \in S_D$ (unit sphere), by f_v we denote the projection of the density f on the line spanned by v.

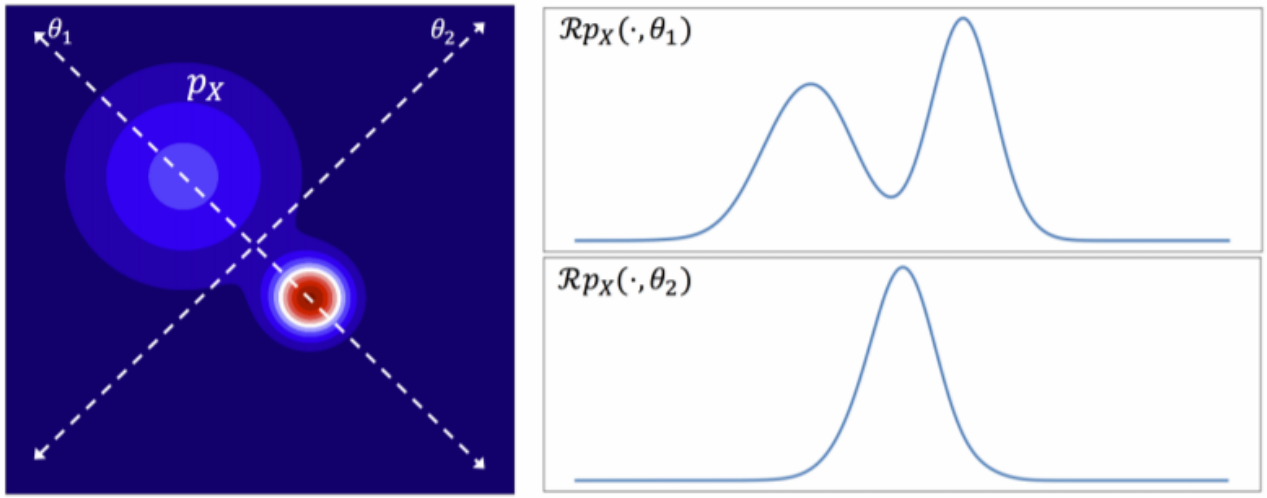


Figure: Sliced model.

Cramer-Wold distance

group of machine
gmum
learning research

By S_D we denote the sphere centered at zero and radius 1 in $\mathbb{R}^D$, and by σ_D denote the normalized surface area measure on $\mathbb{R}^D$, Making use of Cramer-Wold theorem we define the Cramer-Wold distance as the sliced BHEP distance:

$$d^2_{\mathrm{cw}}(f,g) := \int_{S_D} d^2_{\gamma}(f_v, g_v) d\sigma_D(v).$$

Cramer-Wold normality index

group of machine gmum learning research

It occurs that Cramer-Wold distance has a closed form for the distance between spherical gaussian distributions:

$$d^2_{\mathrm{cw}}(\mathcal{N}(x,\alpha I),\mathcal{N}(y,\beta I)) = \frac{1}{\sqrt{2\pi(\alpha+\beta+2\gamma)}}\,{}_1\mathrm{F}_1\left(\tfrac{1}{2};\tfrac{D}{2};\tfrac{\|x-y\|^2}{2(\alpha+\beta+2\gamma)}\right).$$

where ${}_1\mathrm{F}_1$ is the Kummer hypergeometric function.

Making use of this one can easily obtain the formula of the distance of a sample from normal distribution. Consequently, we define the normality index as the normalized distance

$$\mathrm{cw}_D(Z) := \frac{1}{\|N(0,I)\|^2_{\mathrm{cw}}}\, d^2_{\mathrm{cw}}\left(\frac{1}{n}\sum_{i=1}^{n}\delta_{x_i}, N(0,I)\right),$$

where following Bowman-Foster normality test we choose the smoothing parameter γ according to the Silverman's rule of thumb: $\gamma = h^2_{opt}$ where $h_{opt} = \left(\frac{4}{3N}\right)^{1/5}$

Cramer-Wold AutoEncoder (CWAE)

group of machine
gmum
learning research

To ensure that the data transported to latent space $\mathcal{Z}$ are distributed according to the standard normal density, we need to take advantage of the normality index $\mathrm{cw}_D(\mathcal{E}X)$. To obtain a model independent of the possible rescaling of the data, instead of additive, we have decided to use the multiplicative model:

$$\mathrm{cost}(X;\mathcal{E},D) = \mathrm{cw}_D(\mathcal{E}X) \cdot \mathrm{rec\_error}(X;\mathcal{E},\mathcal{D}). \tag{2}$$

Experiments

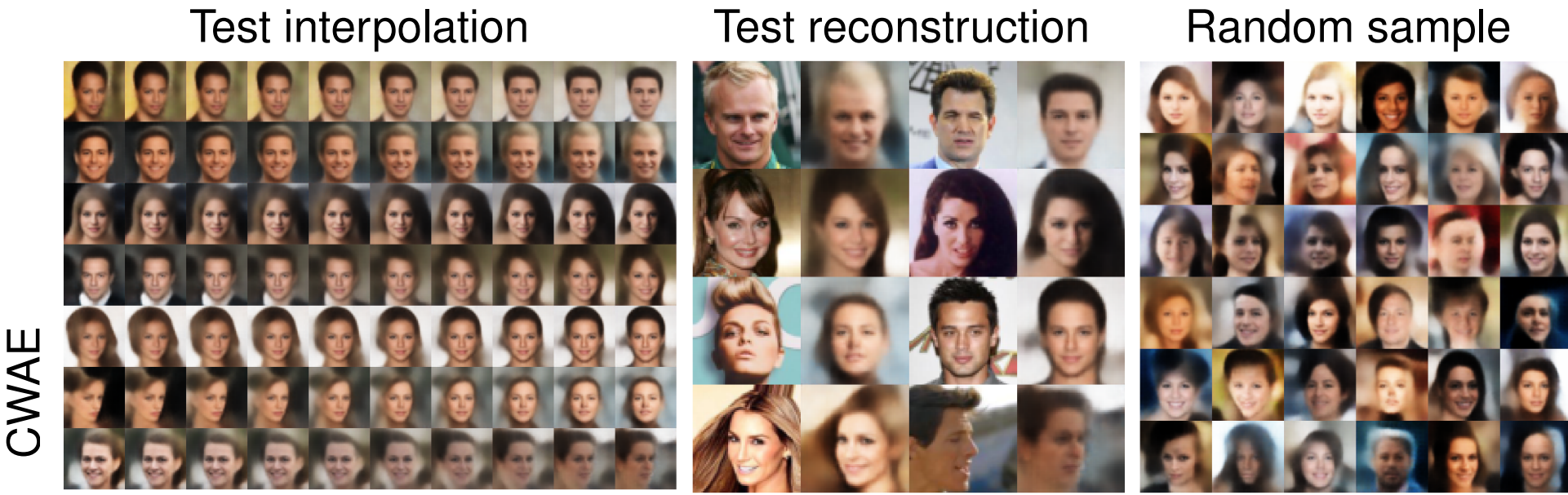


Figure: CWAE on CelebA dataset. In “test reconstructions” odd rows correspond to the real test points.

Experiments

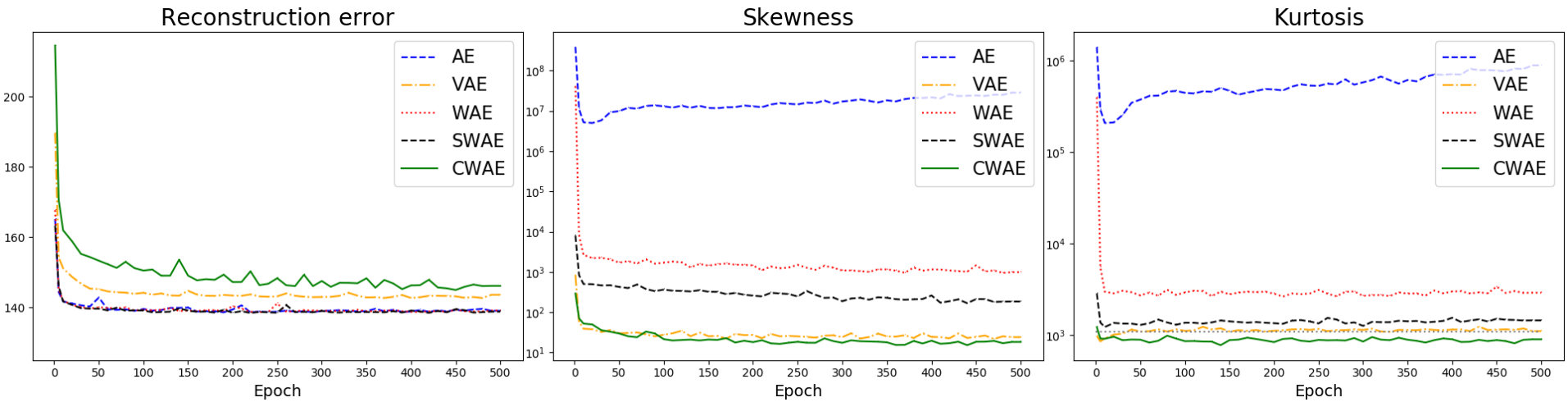


Figure: Value of reconstruction error, Mardia's skewness and normalized kurtosis during learning process of AE, VAE, WAE, SWAE and CWAE on validation dataset in the case of CELEB A datasets. In the case of kurtosis the optimal value is given by the dotted line which denotes the expected value of curtosis for the normal density.

Experiments 2D

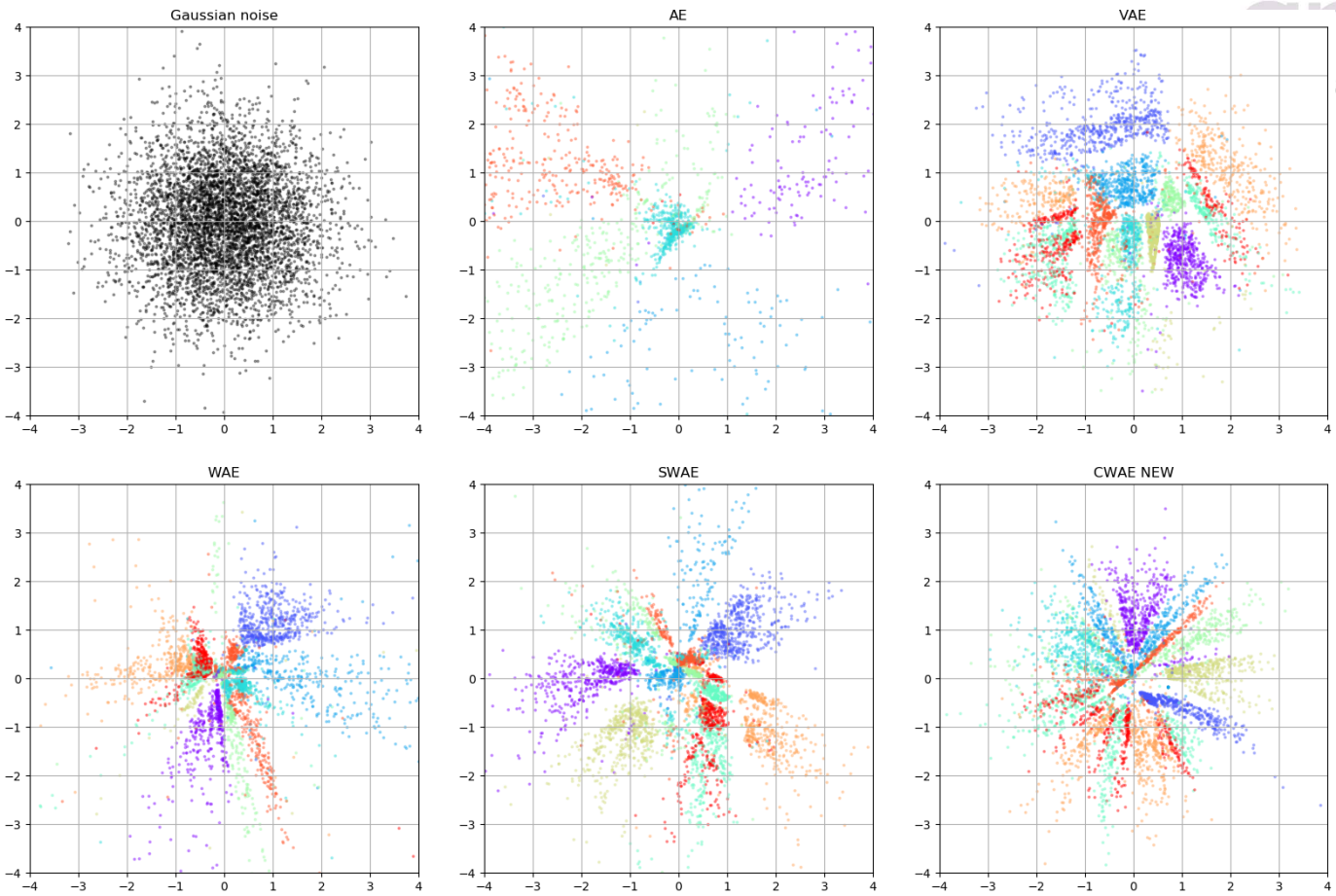


Figure: Two-dimensional latent spaces for AE, VAE, WAE, SWAE, and CWAE, all on MNIST dataset. Models closer to Gaussian noise in the latent space are more generative.

group of machine
gmum
learning research

Norbert Henze.

Invariant tests for multivariate normality: a critical review.

Statistical Papers, 43(4):467–506, 2002.

Bernard W Silverman.

Density estimation for statistics and data analysis, volume 26.

CRC press, 1986.