Variational AutoEncoders with Applications to Anomaly Detection

Luigi Malagò

Transylvanian Institute of Neuroscience (TINS)
Cluj-Napoca, Romania

DEIB, Politecnico di Milano

25 February 2022

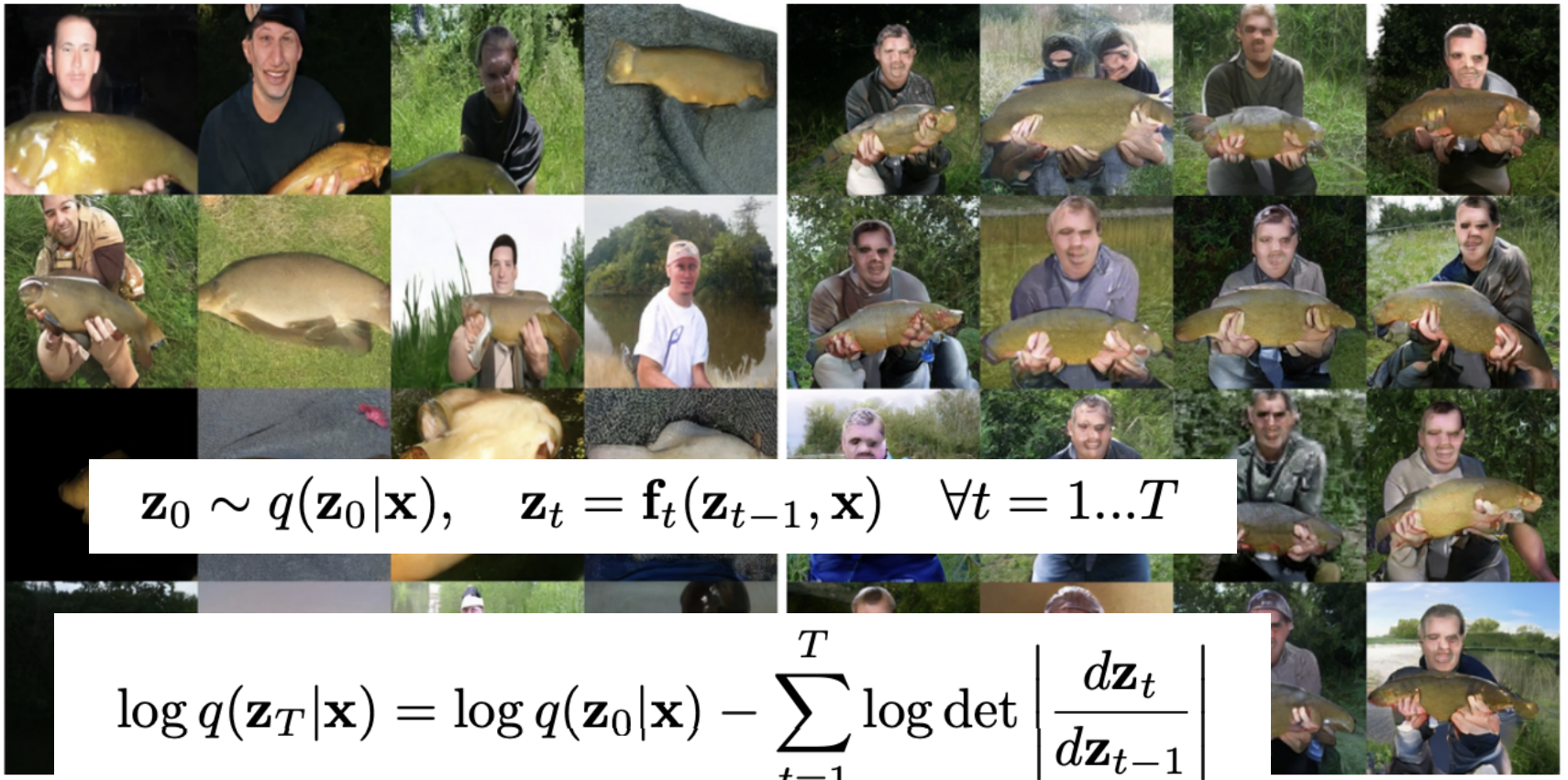
Part 1

Variational AutoEncoders

Plan of the Lecture

- Brief introduction to Variational Inference
- Evidence Lower Bound
- Variational Auto-Encoders
- Geometry of the Latent Space
- Advances in VAEs

Generative Models in Deep Learning



$$\mathbf{z}_0 \sim q(\mathbf{z}_0|\mathbf{x}), \quad \mathbf{z}_t = \mathbf{f}_t(\mathbf{z}_{t-1}, \mathbf{x}) \quad \forall t = 1...T$$

$$\log q(\mathbf{z}_T|\mathbf{x}) = \log q(\mathbf{z}_0|\mathbf{x}) - \sum_{t=1}^{T} \log \det \left| \frac{d\mathbf{z}_t}{d\mathbf{z}_{t-1}} \right|$$

Notation for Bayesian Inference

X, Z multivariate random variables, Z continuous, with probability density functions (pdf) $p(x)$ and $p(z)$ respectively

$p(z)$ is the *prior* and $p(x)$ the *marginal*

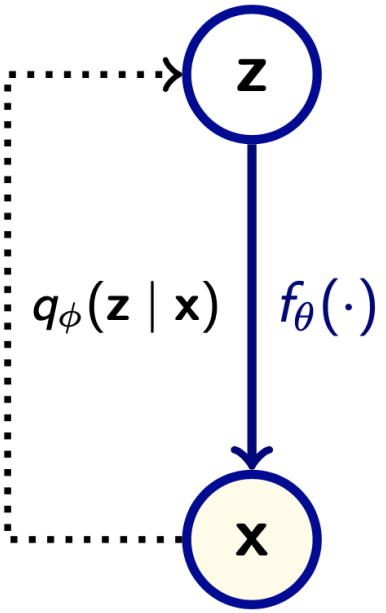
$p(x,z)$ is the pdf of the *joint* random variable (X,Z)

$p(x|z)$, $p(z|x)$ are the conditional pdfs of the random variables $X|Z = z$ and $Z|X = x$

$p(z|x)$ is the posterior

General Setting

The continuous latent r.v. Z generates X, through $f_\theta(\cdot)$ a differentiable function such that $\int p_\theta(x|z)\, p(z)\, dz$ is intractable



The goal is inference, i.e., finding $p_\theta(z|x)$

In Variational Inference [1] we approximate the true posterior $p_\theta(z|x)$ with $q_\varphi(z|x)$, by minimizing the Kullback-Leibler divergence $KL(q_\varphi(z|x)\ \zeta\, p_\theta(z|x))$.

Approach to the solution: maximizing a lower bound of the log likelihood.

Variational Inference 1/2

Deriving the lower-bound

$$\ln p_\theta(\mathbf{x}) = \ln \int q_\phi(\mathbf{z}|\mathbf{x}) \frac{p_\theta(\mathbf{x},\mathbf{z})}{q_\phi(\mathbf{z}|\mathbf{x})} \mathbf{dz} \geq \int q_\phi(\mathbf{z}|\mathbf{x}) \ln \frac{p_\theta(\mathbf{x},\mathbf{z})}{q_\phi(\mathbf{z}|\mathbf{x})} \mathbf{dz} \text{ (Jensen's inequality)}$$

$$\textbf{Evidence lower bound: } \mathcal{L}(\theta,\phi;\mathbf{x}) := \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\ln \frac{p_\theta(\mathbf{x},\mathbf{z})}{q_\phi(\mathbf{z}|\mathbf{x})}\right] \leq \ln p_\theta(\mathbf{x})$$

Minimizing the KL is equivalent to maximizing the lower-bound

$$\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\ln \frac{p_\theta(\mathbf{x},\mathbf{z})}{q_\phi(\mathbf{z}|\mathbf{x})}\right] = \ln p_\theta(\mathbf{x}) - \mathrm{KL}(q_\phi(\mathbf{z}|\mathbf{x}) \parallel p_\theta(\mathbf{z}|\mathbf{x}))$$

The maximum of the lower-bound is the log-likelihood, and it is obtained when $KL(q_\varphi(z|x)\ \zeta\, p_\theta(z|x)) = 0$, thus, the problems are equivalent

Variational Inference 2/2

Optimizing the lower bound maximizes the log likelihood

The distribution of X can be approximated with importance sampling

$$\ln p_\theta(\mathbf{x}) \approx \ln \frac{1}{S} \sum_{i=1}^{S} \frac{p_\theta(\mathbf{x}|\mathbf{z^{(i)}}) p_\theta(\mathbf{z^{(i)}})}{q_\phi(\mathbf{z^{(i)}}|\mathbf{x})}$$

where $z^{(i)} \acute{I} q_\varphi(\cdot|x)$

Fixing the family of distributions for the r.v., e.g. we assume they are Gaussians, we move from variational calculus to regular optimization of the parameters. The problem becomes

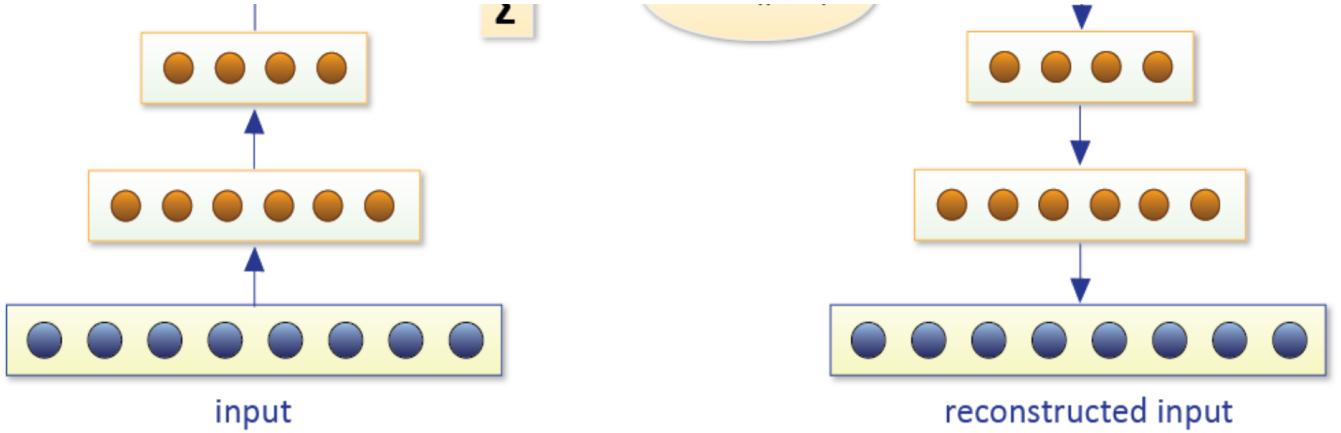
$$\max_{\theta,\phi} \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[\ln p_\theta(\mathbf{x},\mathbf{z}) - \ln q_\phi(\mathbf{z}|\mathbf{x})]$$

Variational AutoEncoders

Variational AutoEncoders [6], [11] tackle the problem of variational inference in the context of neural networks

The parameters φ and θ of $q_\varphi(z|x)$ and $p_\theta(x|z)$ are learned through two different neural networks: the *encoder* and the *decoder*

$$\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\ln \frac{p_\theta(\mathbf{x},\mathbf{z})}{q_\phi(\mathbf{z}|\mathbf{x})}\right] = \ln p_\theta(\mathbf{x}) - \mathrm{KL}(q_\phi(\mathbf{z}|\mathbf{x}) \parallel p_\theta(\mathbf{z}|\mathbf{x}))$$



Applications

- Encode data point: learn a lower dimensional representation of the dataset, by sampling from $q_\varphi(\cdot|x)$
- The dimension of the latent variable Z is assumed to be much smaller than the dimension of the dataset.
-
- Generate new samples from noise examples that resemble the ones seen during training
- The prior $p(z)$ on the latent variable is assumed Gaussian $N(0,I)$ and samples are fed through the network to output the conditional probabilities $p_\theta(x \mid z)$

Details of the Algorithm

Encoder $q_\varphi(z|x)$ - Gaussian *N(μ, D)* with diagonal covariance, parametrized by φ

Decoder $p_\theta(x|z)$ - Gaussian with diagonal covariance (continuous data) or Bernoulli vector (discrete data) parametrized by θ i

For a data point x, rewrite the lower bound *L(θ, φ; x)*

$$\mathcal{L}(\theta, \phi; \mathbf{x}) = \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[\ln p_\theta(\mathbf{x}|\mathbf{z})] - KL(q_\phi(\mathbf{z}|\mathbf{x}) \,||\, p_\theta(\mathbf{z}))$$

Reconstruction error — Regularization

Cost function to be optimized: $\frac{1}{N}\sum_{n=1}^{N}\mathcal{L}(\theta, \phi; \mathbf{x}_n)$, from dataset $\mathbf{X} = \{\mathbf{x}_n\}_{n=\overline{1,N}}$

Back-propagating through Stochastic Layers

Training neural networks requires computing the gradient of the cost function, using back-propagation

Difficulty when computing $\nabla_{\varphi} E_{q\varphi(z|x)}[\ln p_{\theta}(x|z)]$, indeed the Monte Carlo estimation of the gradient has high variance

The reparameterization trick: find $g_{\varphi}(\cdot)$ differentiable transformation and random variable Γ with pdf $p(\cdot)$, such that $Z = g_{\varphi}(\Gamma)$

$$\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[\ln p_\theta(\mathbf{x}|\mathbf{z})] = \mathbb{E}_{p(\gamma)}[\ln p_\theta(x|g_\phi(\gamma))]$$

$$\nabla_\phi \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[\ln p_\theta(\mathbf{x}|\mathbf{z})] = \mathbb{E}_{p(\gamma)}[\nabla_\phi \ln p_\theta(x|g_\phi(\gamma)]$$

Example: for $X \acute{I} N(\mu,\Sigma)$, with $\Sigma = LL^T$ Cholesky decomposition

$X = \mu + L\Gamma$, with $\Gamma \acute{I} N(0, I)$

Limitations and Challenges

Limitations

- The conditional independence assumption on the latent variables given the observations limits the expressive power of the approximate posterior
- Limitation on the number of active latent variables when using a hierarchy of stochastic layers [13]

Challenges

- Difficulty when training on text data: empirical observation that the learned latent representation is not meaningful [2]
- How to improve the quality of the generated samples, in case of a dataset of images? How can we find a better correlation between the images generated and the maximization of the lower bound?
- How to estimate the tightness of the bound?

Research Directions

More complex representations for $q_\varphi(z|x)$, by transforming a simple distribution through invertible differentiable functions, as in [10] and [5]

Increased complexity of the graphical models, e.g. a hierarchy of latent variables or auxiliary variables as in [13] and [9]

Designing tighter bounds

- importance weighting estimates of the log-likelihood [3]

$$\mathcal{L}_K(\phi, \theta; \mathbf{x}) = \mathbb{E}_{\mathbf{z}^1, \mathbf{z}^2, \ldots \mathbf{z}^K \sim q_\phi(\mathbf{z}|\mathbf{x})} \left[\log \frac{1}{K} \sum_{k=1}^{K} \frac{p_\theta(\mathbf{x}, \mathbf{z}^k)}{q_\phi(\mathbf{z}^k|\mathbf{x})} \right]$$

- minimizing different divergences (Renyi [8] and α-divergence [4])

Overcoming the challenge of training VAE on text data [2]

Gaussian Graphical Models for the Latent Variables

- Gaussian Graphical Models [7] introduce correlations in the latent variables
- For instance a *chain* of r.v. sparse precision matrix $P = \Sigma^{-1}$ with number of non-zero components linear in the dimension of the latent variable
- The encoder network outputs the mean μ and the Cholesky factor L of the precision matrix. L will have a special sparse structure and will ensure the positive definiteness of Σ.
- To sample from $N(\mu, \Sigma)$: solve linear system $L^T v = \varepsilon$, where $\varepsilon \acute{I} N(0, I)$, and output $z = \mu + v$
- Sampling from $N(\mu, \Sigma)$ and computing $KL(N(\mu, \Sigma) \zeta N(0, I))$ can be done in linear time introduce expressiveness without extra computational complexity.

Chain of Gaussian Random Variables

Chain Model

$$P = \begin{pmatrix} \sigma_1 & \lambda_1 & & 0 \\ \lambda_1 & \sigma_2 & \lambda_2 & \\ & & \ddots & \\ 0 & & \lambda_{k-1} & \sigma_k \end{pmatrix}$$

The precision matrix P is tridiagonal

The Cholesky factor of such a matrix is lower-bidiagonal

Analysis of the Representations in the Latent Space

Experiments on MNIST dataset to understand the representation of the images in the learned latent space

Principal Components Analysis of the latent means will give us insights about which components are relevant for the representation.

The components with a low variation along the dataset are the ones not meaningful.

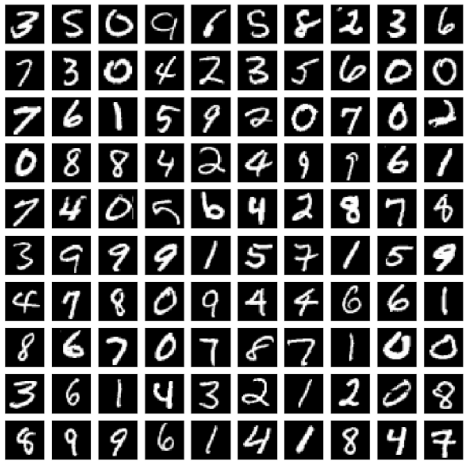
PCA eigenvalues of the posterior samples are very close to 1
 the KL minimization forces some components to be $N(0, I)$

Interpretation of the Plot

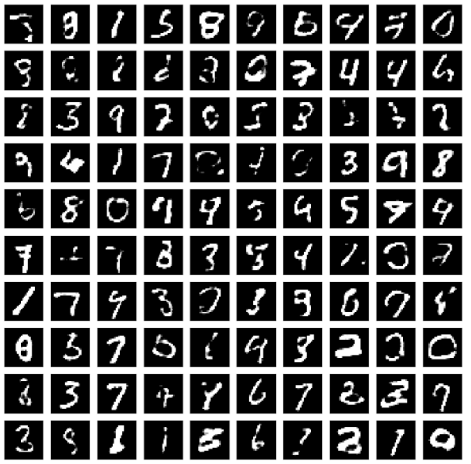
VAE trained with latent size 20 on MNIST, only around 15 of the latent variables are relevant for the representation

The number remains constant when training with a larger latent size

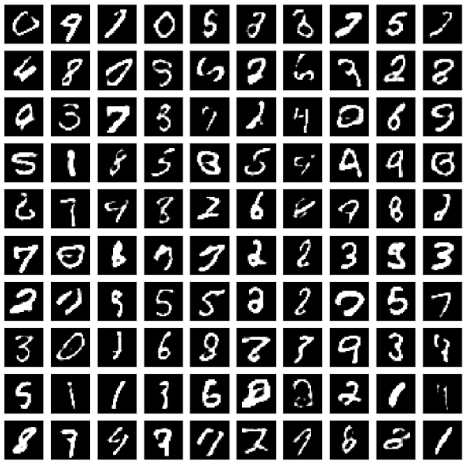
This is a consequence of the KL regularization term in the ELBO, which forces some components to be Gaussian noise



Samples from MNIST dataset

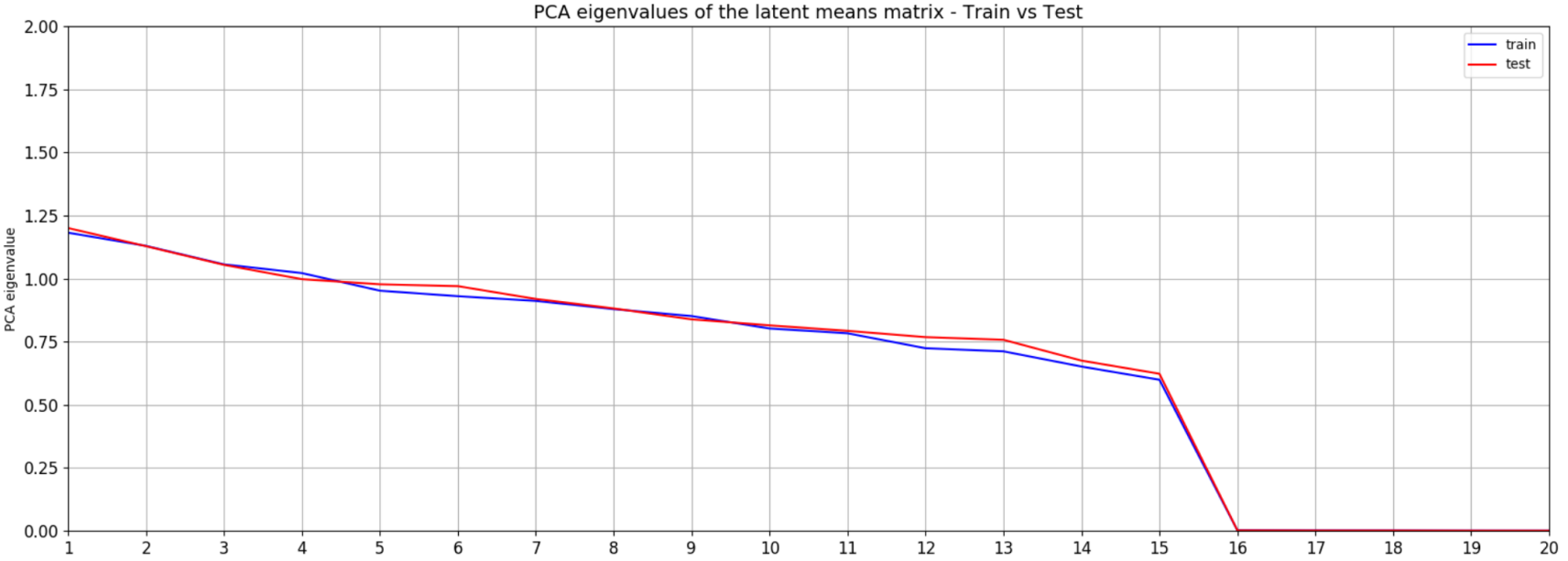


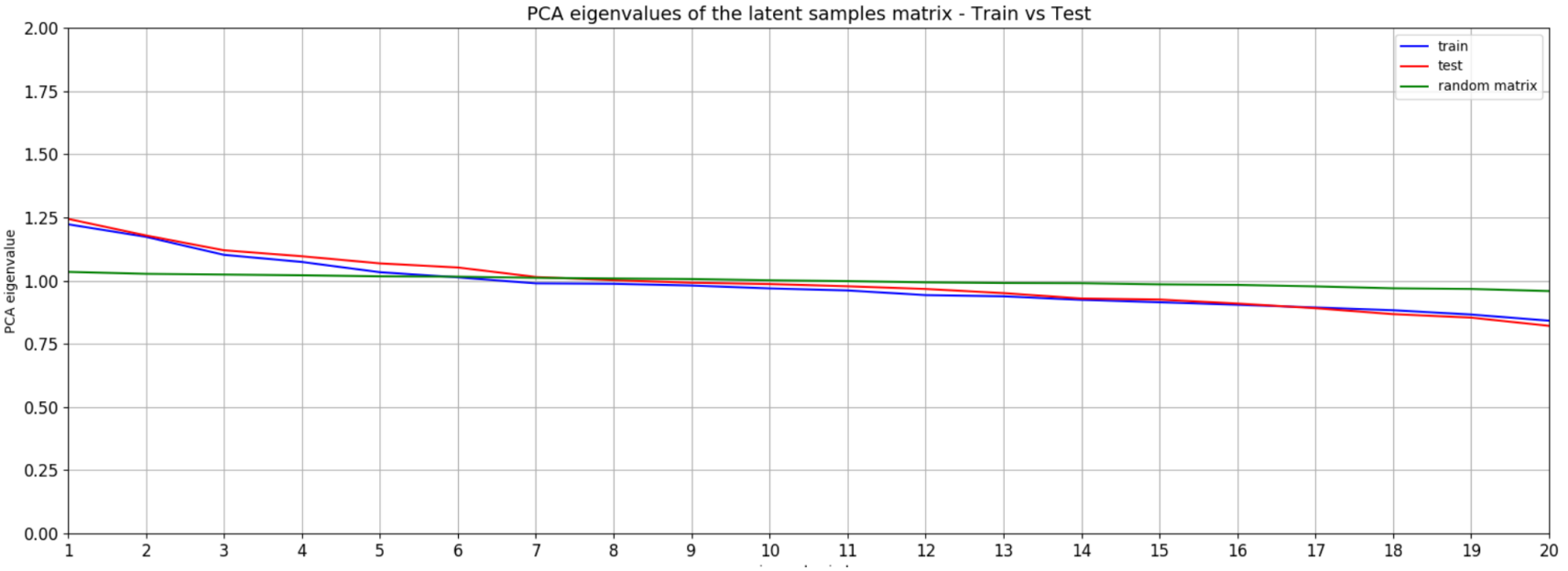
Generated after 100 epochs



Generated after 1000 epochs

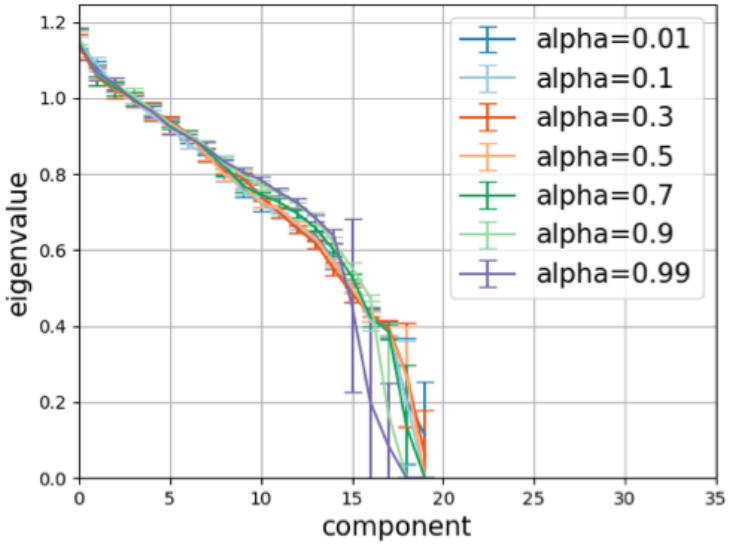
PCA Plots


PCA eigenvalues of the latent means matrix - Train vs Test
PCA eigenvalue
train
test
2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20

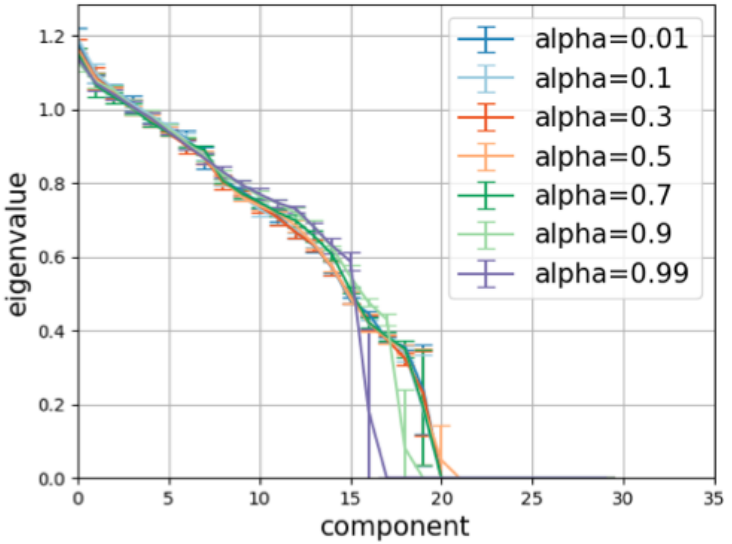

PCA eigenvalues of the latent samples matrix - Train vs Test
PCA eigenvalue
train
test
random matrix
2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20

Identification of variables carring information

$k = 20$, RELU



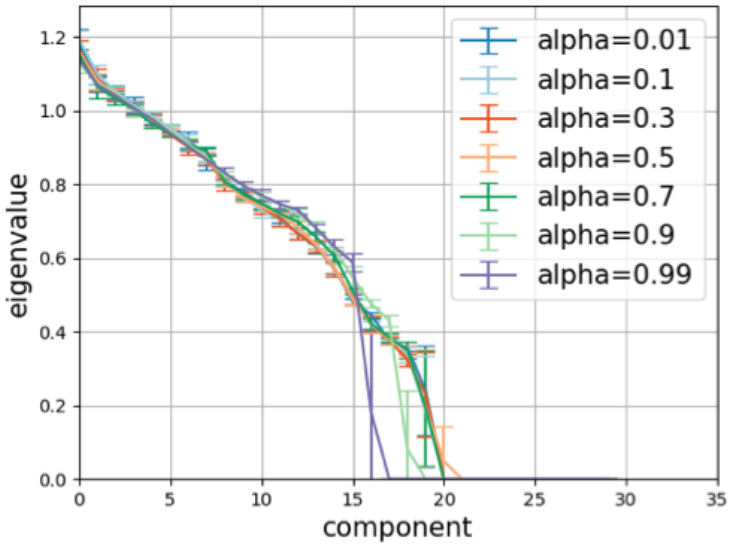
$k = 30$, RELU



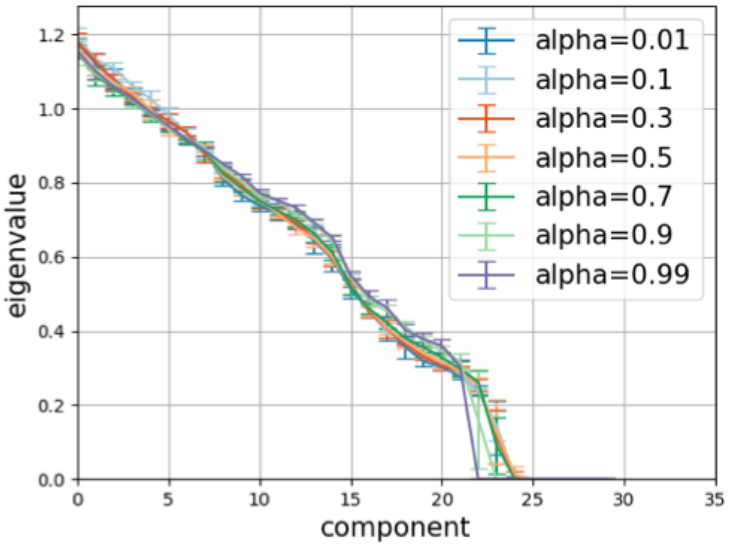
The number of components carrying information slight change with the type of lower bound, but not with the hidden space dimension

Identification of variables carrying information

$k = 30$, RELU



$k = 30$, ELU



The number of components carrying information slight change with the type of lower bound, but not with the hidden space dimension

Sparse representations / data compression

We observe, sparsity which implies data compression

The level of sparsity is

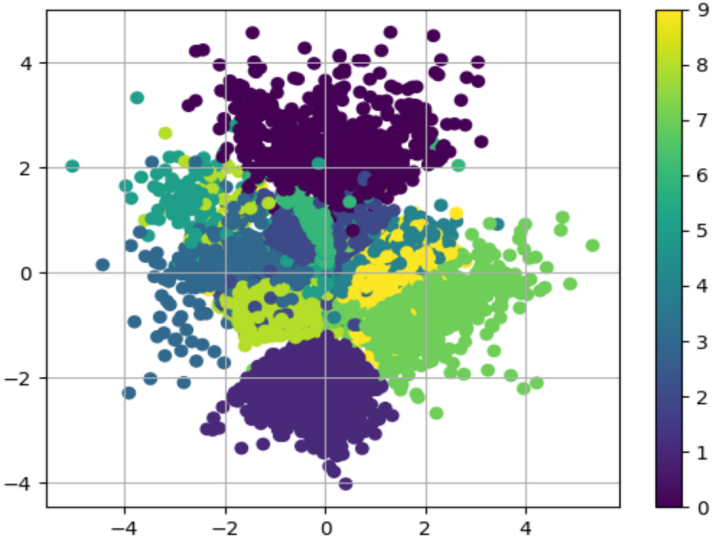
- controlled by the regularization parameter
- independent from the latent space dimension (for dimension large enough)
- independent from the encoder network topology (for networks which are large enough)

Sparsity appears also for generalized lower bounds (for instance based on Renyi divergence)

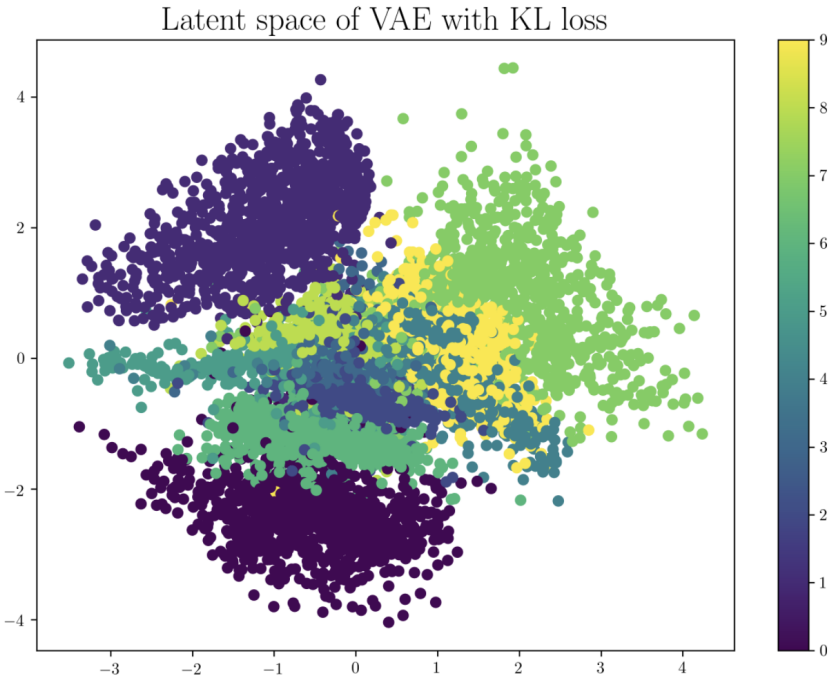
Geometry of the Latent Space: VAE vs AE

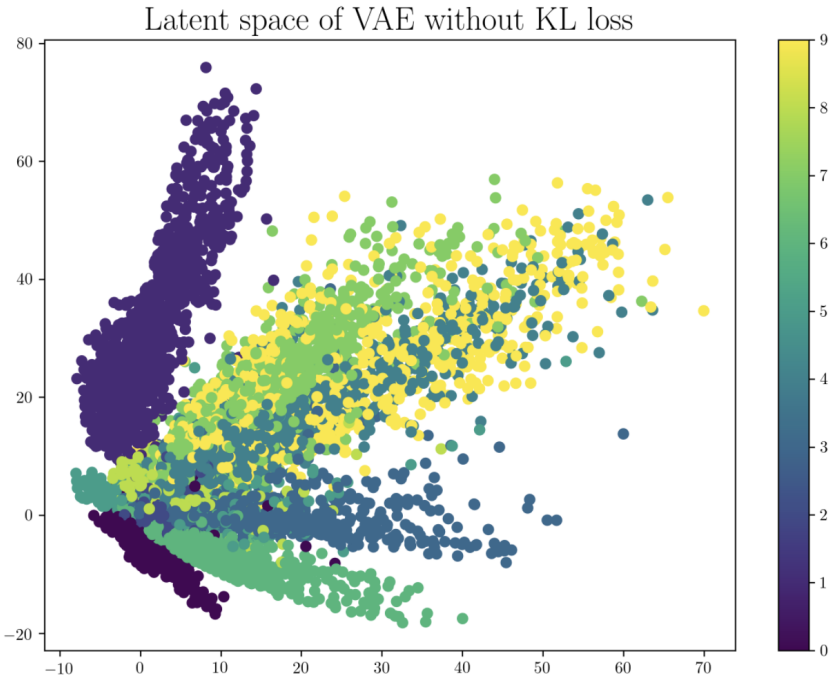
VAE is trained on MNIST with 2 latent variables. The plot represents the means of the posterior for each point in the dataset, colored by corresponding class

- Linear separability of the classes in the space of latent representations
- Sampling in the latent space from the empty regions images that are not digits
- Linear interpolation property continuous deformation in the latent space between two different images



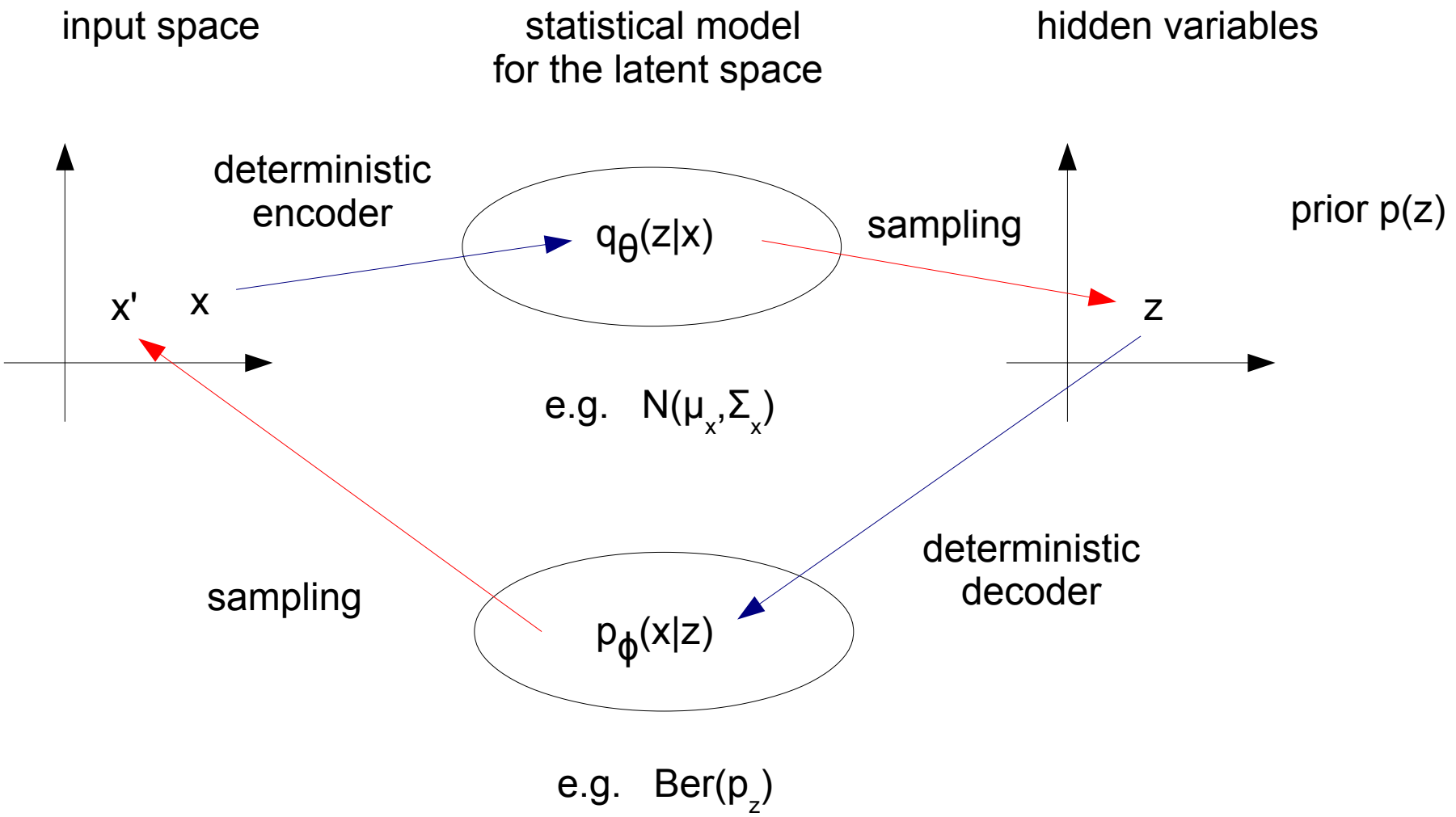
Geometry of the Latent Space: VAE vs AE



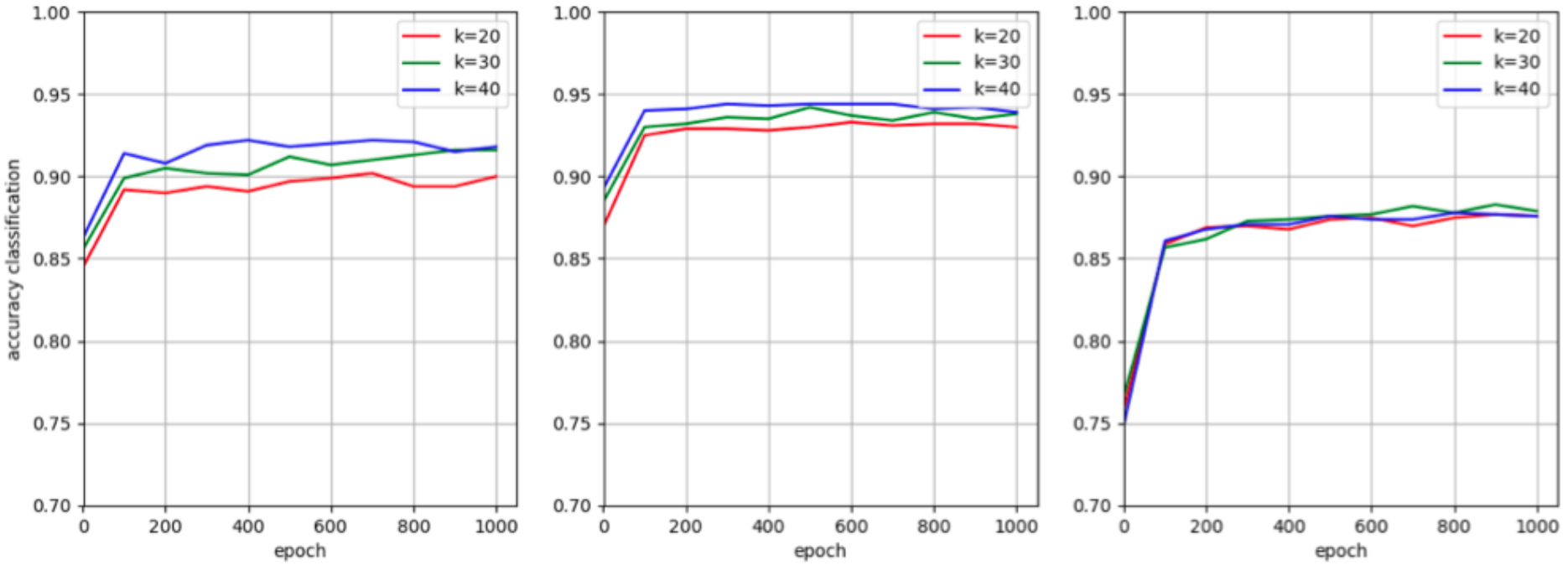


Source: Difference between AutoEncoder (AE) and Variational AutoEncoder (VAE), Aqeel Anwar, https://towardsdatascience.com/

Maps from Observations to Statistical Models

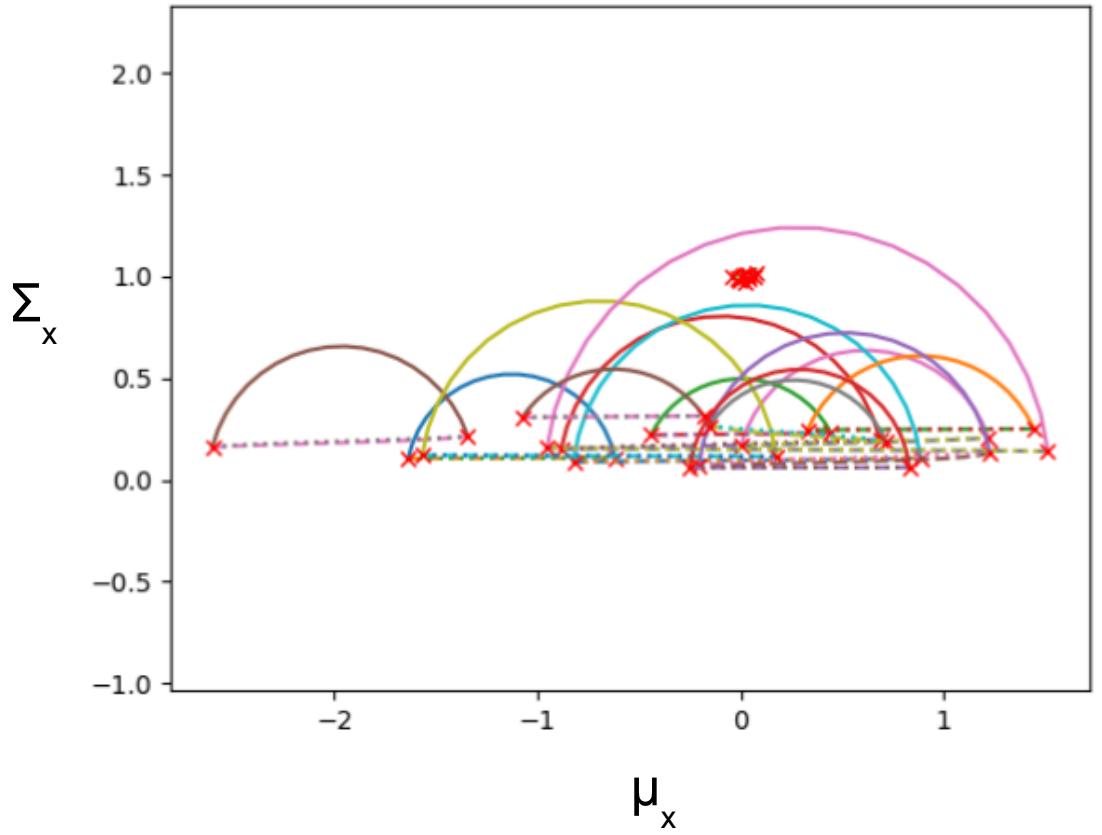


Linear separability over the manifold $q_\theta(z|x)$



The manifold parameterization (μ_x, Σ_x) carries more structure than the latent space

Geodesics over independent Gaussians



Statistical models admit a **Riemannian geometry** based on the Fisher Rao metric, which gives invariant notions of distances, geodetics, gradients (Amari, 1982)

Interpolations of Images

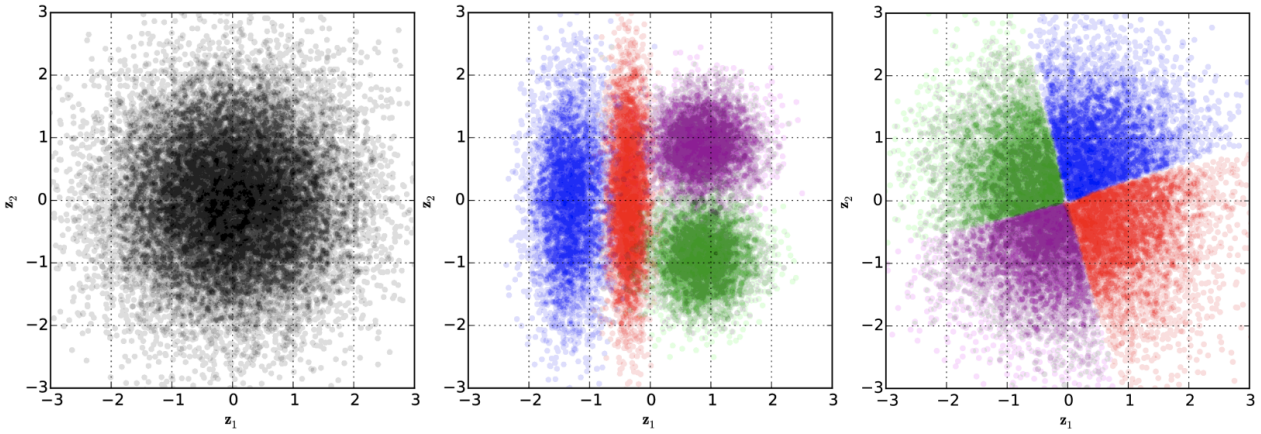
- Linear interpolation in the input space
- Linear interpolation between the latent mean vectors
- Linear interpolation between two hidden representations
- Decoded samples from the normal distributions found along the Fisher-Rao geodesic in the latent space
- Decoded latent means of the normal distributions found along the Fisher-Rao geodesic in the latent space

Advances in VAEs: Normalizing Flows

Normalizing Flow (Rezende et al., 2015) and IAF (Kingma et. al, 2016) allow to build building flexible posterior distributions through an iterative procedure

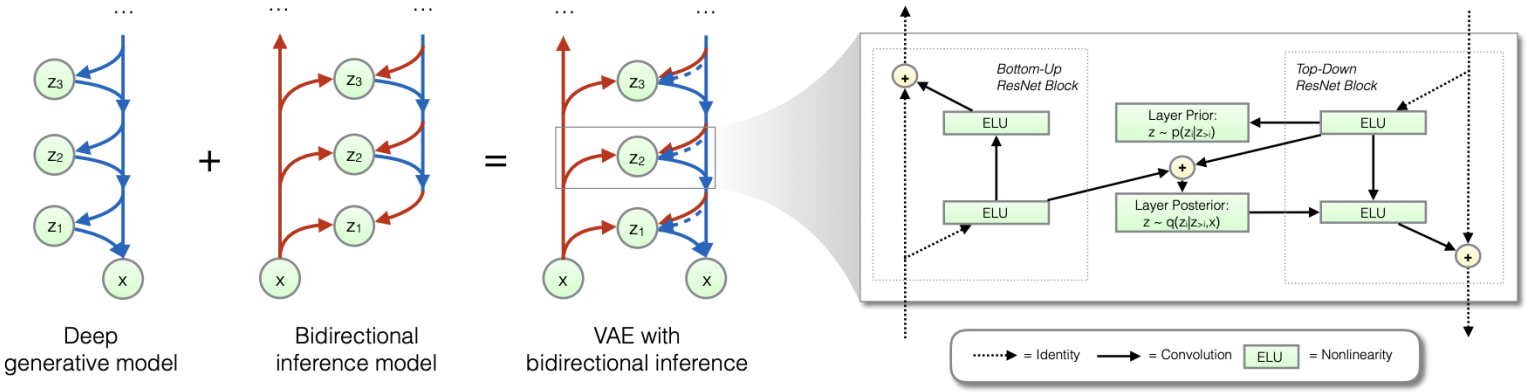
$$\mathbf{z}_0 \sim q(\mathbf{z}_0|\mathbf{x}), \quad \mathbf{z}_t = \mathbf{f}_t(\mathbf{z}_{t-1}, \mathbf{x}) \quad \forall t = 1...T$$

$$\log q(\mathbf{z}_T|\mathbf{x}) = \log q(\mathbf{z}_0|\mathbf{x}) - \sum_{t=1}^{T} \log \det \left| \frac{d\mathbf{z}_t}{d\mathbf{z}_{t-1}} \right|$$



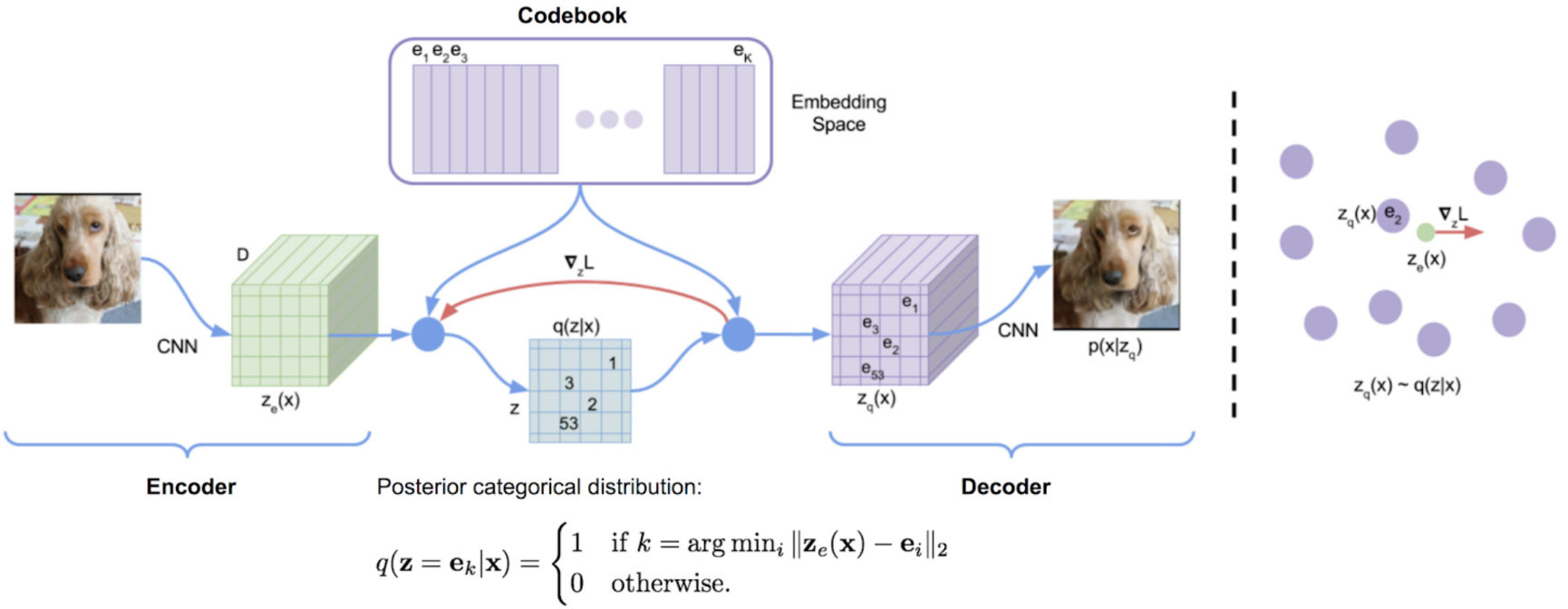
(a) Prior distribution (b) Posteriors in standard VAE (c) Posteriors in VAE with IAF

Advances in VAEs: Bidirectional Inference



Inverse Autoregressive Flows (Kingma et. al, 2016)

Advances in VAEs: Vector Quantization



$$q(\mathbf{z} = \mathbf{e}_k|\mathbf{x}) = \begin{cases} 1 & \text{if } k = \arg\min_i \|\mathbf{z}_e(\mathbf{x}) - \mathbf{e}_i\|_2 \\ 0 & \text{otherwise.} \end{cases}$$

VQ-VAE (Aaron van den Oord et al., 2018)

- Discrete representation learned during training
- Nearest element embedding
- No backpropagation, gradient is estimate as in straight-through estimator (from decoder input to encoder output)

References 1/2

[1] David M Blei, Alp Kucukelbir, and Jon D McAuliffe. Variational inference: A review for statisticians. Journal of the American Statistical Association, 2017.

[2] Samuel R Bowman, Luke Vilnis, Oriol Vinyals, Andrew M Dai, Rafal Jozefowicz, and Samy Bengio. Generating sentences from a continuous space. arXiv preprint arXiv:1511.06349, 2015.

[3] Yuri Burda, Roger Grosse, and Ruslan Salakhutdinov. Importance weighted autoencoders. arXiv preprint arXiv:1509.00519, 2015.

[4] Jos´e Miguel Hern´andez-Lobato, Yingzhen Li, Mark Rowland, Daniel Hern´andez-Lobato, Thang D Bui, and Richard E Turner. Black-box α-divergence minimization. 2016.

[5] Diederik P Kingma, Tim Salimans, Rafal Jozefowicz, Xi Chen, Ilya Sutskever, and Max Welling. Improving variational autoencoders with inverse autoregressive flow. In Advances In Neural Information Processing Systems, pages 4736–4744, 2016.

[6] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. 2013.

[7] Steffen L Lauritzen. Graphical models, volume 17. Clarendon Press, 1996.

[8] Yingzhen Li and Richard E Turner. R´enyi divergence variational inference. In Advances in Neural Information Processing Systems, pages 1073–1081, 2016.

References 2/2

[9] Lars Maaløe, Casper Kaae Sønderby, Søren Kaae Sønderby, and Ole Winther. Auxiliary deep generative models. arXiv preprint arXiv:1602.05473, 2016.

[10] Danilo Jimenez Rezende and Shakir Mohamed. Variational inference with normalizing f bws. arXiv preprint arXiv:1505.05770, 2015.

[11] Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. 2014.

[12] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. In Advances in Neural Information Processing Systems, pages 2234–2242, 2016.

[13] Casper Kaae Sønderby, Tapani Raiko, Lars Maaløe, Søren Kaae Sønderby, and Ole Winther Ladder variational autoencoders. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, Advances in Neural Information Processing Systems 29, pages 3738–3746. Curran Associates, Inc., 2016.

Part 2a

Anomaly Detection based on Variational AutoEncoders

Surveys

- Guansong Pang, Chunhua Shen, Longbing Cao, Anton van den Hengel, Deep Learning for Anomaly Detection: A Review, arXiv:2007.02500, 2020

- Lukas Ruff, Jacob R. Kauffmann, Robert A. Vandermeulen, Grégoire Montavon, Wojciech Samek, Marius Kloft, Thomas G. Dietterich, Klaus-Robert Müller, A Unifying Review of Deep and Shallow Anomaly Detection, arXiv:2009.11732, 2020

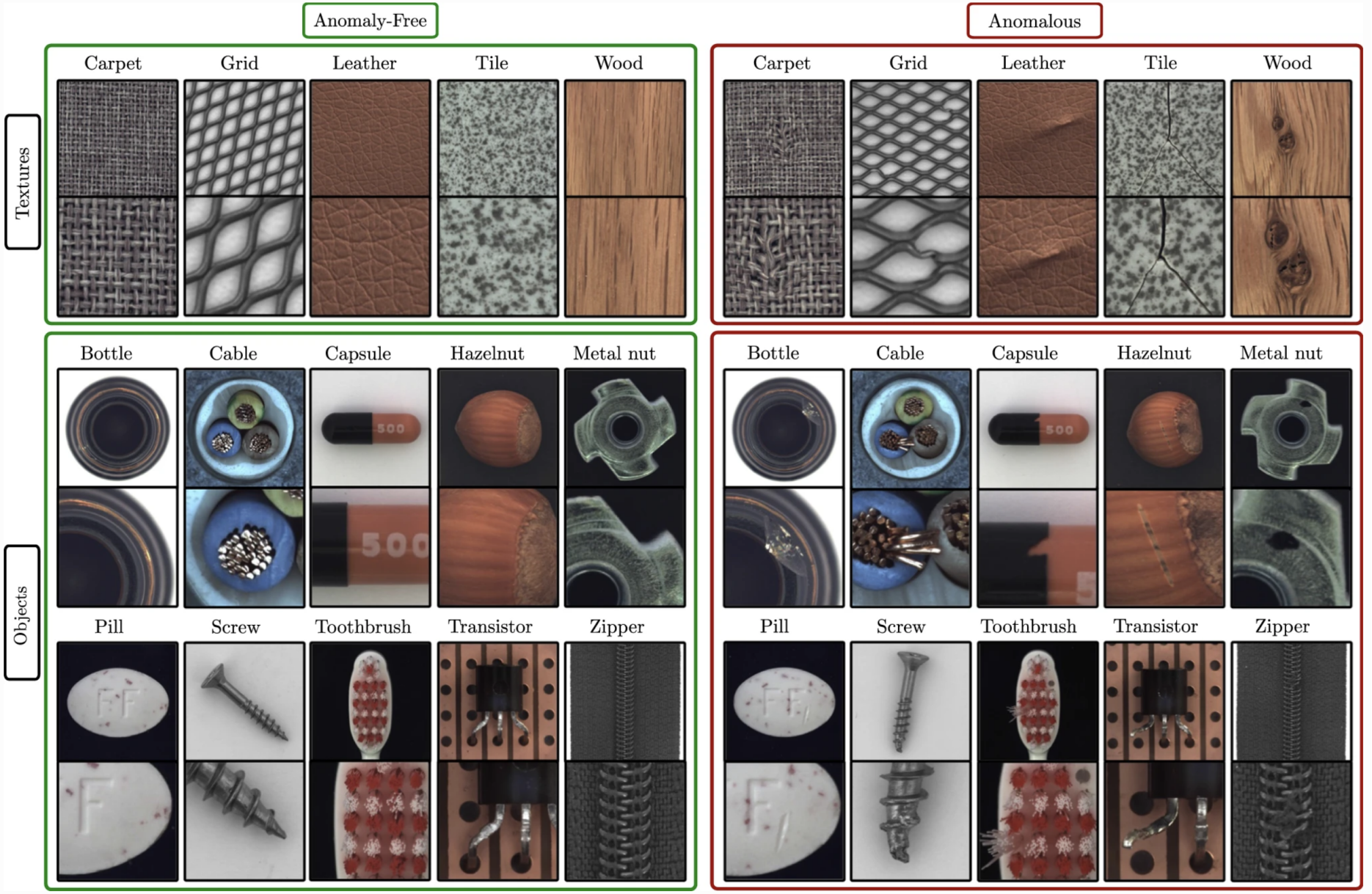
What is Anomaly Detection?

Identification of instances which are not normal/regular w.r.t. a set of observations

Several challenges

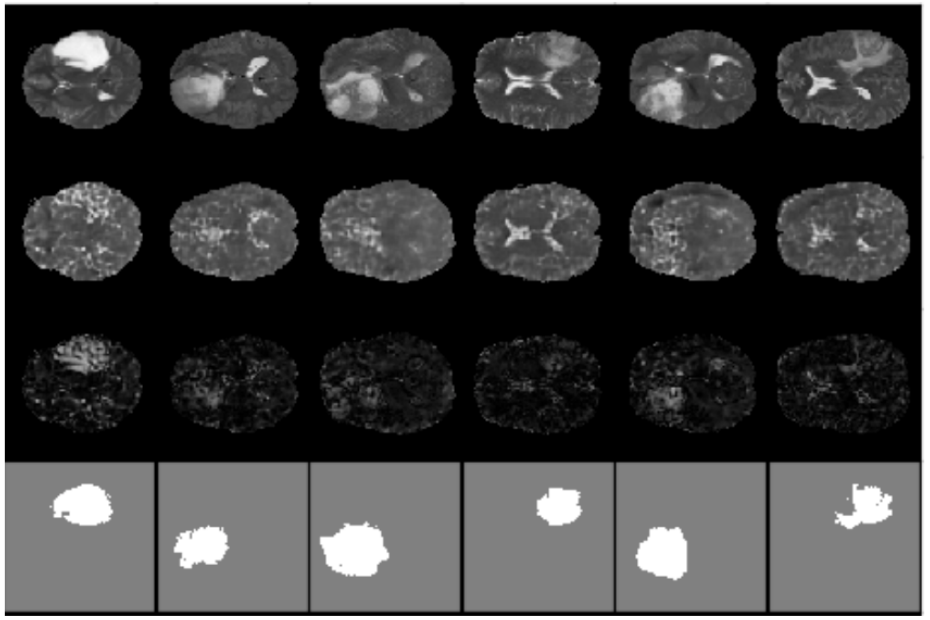
- Anomalies are usually uncommon (few samples, difficulties to label data)
- Anomalies may not be known in advance when the algorithm is designed (or trained)
- Interpretability: Identification of the anomaly in the instance (e.g., portion of image or time window for a signal)
- Different domains: images, signals, categorical data, time series, etc

Examples of Anomalies in Industrial Applications

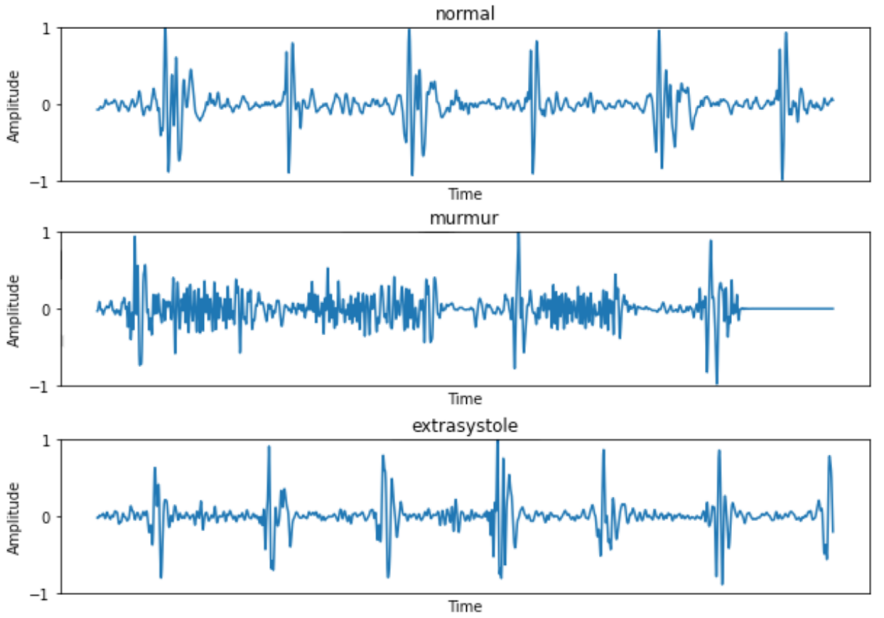


MVTec Anomaly Detection Dataset - MVTec AD

Examples of Anomalies in Healthcare



BRATS dataset



Heartbeat dataset

Many Approaches, No Unified Taxonomy

Supervised vs weakly-supervised vs unsupervised methods

Model-based vs distance-based anomaly detection

Clustering-based (distance are used and a cluster is a model)

Training (requires a training set) vs non-training of a model (data are directly fed to the algorithm)

Methods based on Deep Learning: “Deep” Anomaly Detection

Supervised vs Unsupervised Methods

Supervised methods

- A labelled dataset is available
- Anomaly detection can be formalize as binary classification dataset
- Anomalies are commonly rare: high imbalance

Non-supervised methods

- Anomalies are typically not used in training
- Only regular instances are used in training: weakly/partially or semi-supervised
- Since many methods are robust w.r.t. outliers in training, semi-supervised methods are referred as unsupervised

Distance-based Anomaly Detection

- These methods rely on a distance computation to identify anomalies
- Distances are used to compute densities, anomalies appear in low density regions

Distances are used to compute densities in

- nearest-neighbor-based methods
- density-based methods

Example: *Local Outlier Factor* (LOF)

The density of a data point is compared with those of its k-nearest neighbors under the assumption that for inliers these two quantities are similar

Model-based Anomaly Detection

A model of the regular instances in learned in training (typically in an unsupervised or weakly supervised way)

Outliers are identified as those observations which do not fit the model

Usually robust to outliers when the method is unsupervised

Example: *OneClass Support Vector Machines*

Data points are projected in a high dimensional space using kernels

OneClass SVM computes a boundary around the data points together with a decision function, by solving an opt problem

The score function is based on the margin

Clustering-based Anomaly Detection

The main idea is to cluster data set and then to flag anomalies as those data points which do not belong to any cluster

In unsupervised methods, clustering algorithms need to be robust, to avoid to add outliers to a cluster

If multiple outliers form a cluster, they may be considered as regular points (inspection of small or sparse clusters)

Example is FindOut [13], which does not force outliers into clusters, or k-means with extra checks for anomalous clusters

“Deep” Anomaly Detection

A set of methods for Anomaly Detection based on the use of Deep Learning models

- Features Extractors (e.g., classifications model used as features extractors)
- Reconstruction Models (AE, VAE, etc)
- Generative Models (VAE, GAN, etc)

DL allows an end-to-end approach to Anomaly Detection: the anomaly score is learned in training simultaneously with the features

A Toy Example 1/2

Take a pre-trained model, e.g., VGG16 on ImageNet

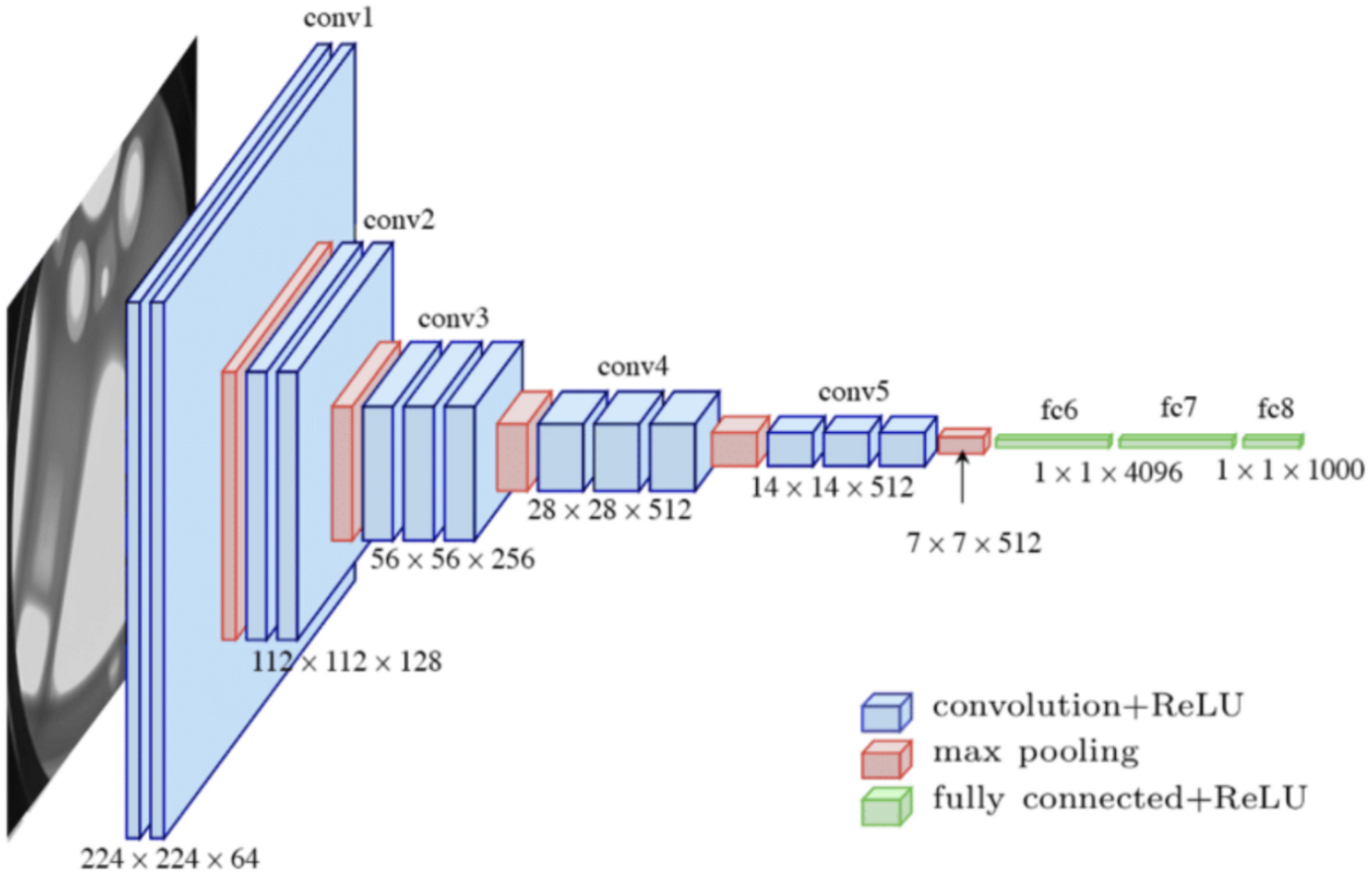
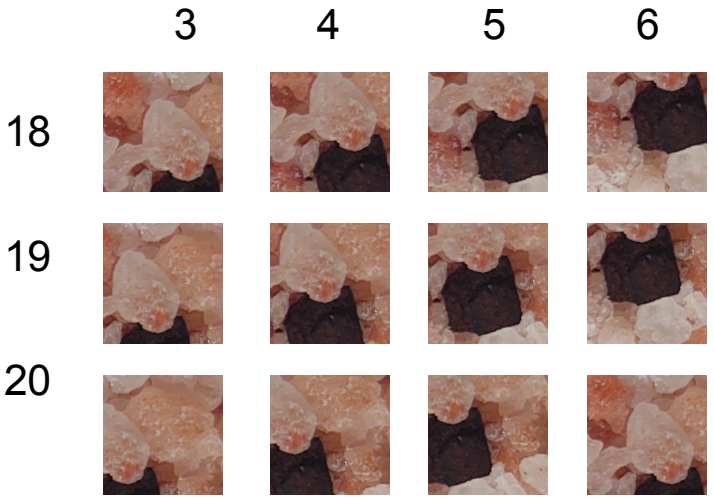
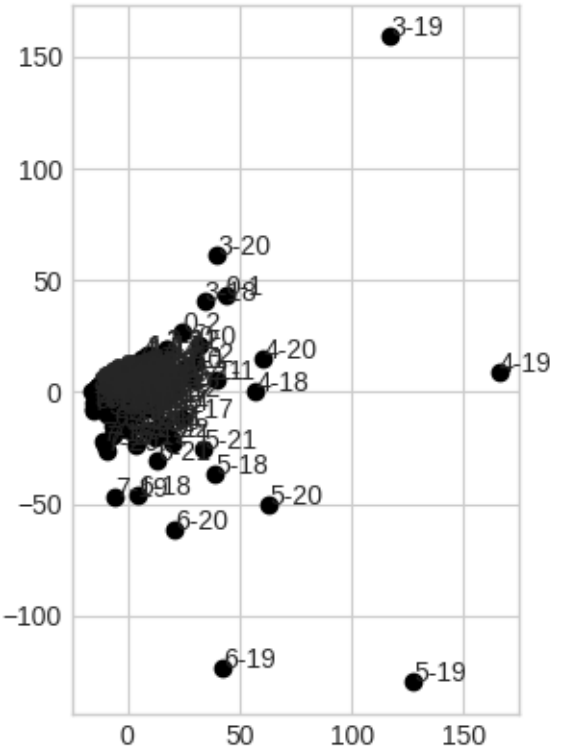


Figure 1: **The architecture of VGG16**. Source: Researchgate.net

A Toy Example 2/2

Features extracted from fc6 layer, followed by a 2-dim PCA





Anomaly Detection using (Variational) AutoEncoders

Main characteristics of AutoEncoders (AE)

- provide a compact hidden representation (dimensionality reduction)
- are trained to reconstruct good quality images

Additionally, Variational AutoEncoders (VAE)

- are generative models with a prior distribution on the latent space
- the bottleneck layer is obtained by sampling from a statistical model (approximate posterior)

Anomaly Scores with (Variational) AutoEncoders

In addition anomaly scores

- can be defined in the *input* space (high-dimensional) or in the *latent* space (compact representations)
- may have a statistical interpretation
  - AE can be trained by maximization of the likelihood
  - VAE are trained my the maximization of a lower-bound of the likelihood, the ELBO)
  - for VAE, anomaly scores can be defined over the space of approximate posteriors

Anomaly Scores for AEs: Input Space

An AE trained on regular instances only should not perform well in reconstructing anomalies

Main idea: the model has learned a truth manifold structure of prototypical representation, and thus large reconstruction errors would imply off-manifold or non-prototypical instances

A candidate reconstruction loss is given by the reconstruction error $s(x) = || x - r \circ e(x)||^2$

Typically, no statistical interpretation for such reconstruction loss

Anomaly Scores for AEs: Latent Space

Distance functions can loose their meaning and function in high dimensions

AEs provide compact representation of the input data in the bottleneck layer

Anomaly Scores typically used with shallow models can be used on the lower-dimensional latent representations

Examples

- Distance-based Anomaly Detection (Local Outlier Factor)
- Model-based Anomaly Detection (OneClass SVM)
- Clustering-based Anomaly Detection (FindOut, k-means)

Typically better performance compare to the input space

Anomaly Scores for VAEs: Reconstruction Loss

Anomaly scores based on $p(x)$

1. Sample z from the prior $p(z)$
2. Estimate $p(x)$ through $E_{z\acute{}p(z)}\,[p_\theta(x|z)]$

This has a nice theoretical interpretation, but it performs worse than conditioning on x

1. Sample z from $q_\phi(z|x)$
2. Estimate $p(x)$ through $E_{z\acute{}q\phi(z|x)}\,[p_\theta(x|z)]$

(probabilistic reconstruction model)

A Partial List of Algorithms based on VAE for AD

- Vuet al. (2019) introduced an approach combining VAEs and adversarial training
- João Pereira et. al (2018) introduce a local metric based on Wasserstein distance in the latent space
- Zimmerer et al. (2019) proposed to combine context-encoders and VAEs to obtain a more robust anomaly detection framework

- Chen et al. (2018) a regularizer that encourages the learning of more suitable latent space representations for unhealthy images and their healthy counterparts was introduced (representation consistency)

  => *they observe that abnormal images are not necessarily mapped outside the predetermined latent distribution*

Identifying Where the Anomaly Appears

Identifying the anomaly in the dataset is a more complex task than simply classifying anomalous vs regular instances

Several alternatives are possible

- If the Anomaly Score is compute in the input space, the segmentation task can be obtained by taking the difference between the input image and its reconstruction
- If the Anomaly Score is computed in the latent space, a directional derivative in the “direction of normality” can be computed with respect to the pixels of the image

Part 2b

Two Applications in Anomaly Detection in Healthcare based on Variational AutoEncoders

Thanks for your attention

www.luigimalago.it

malago@tins.ro