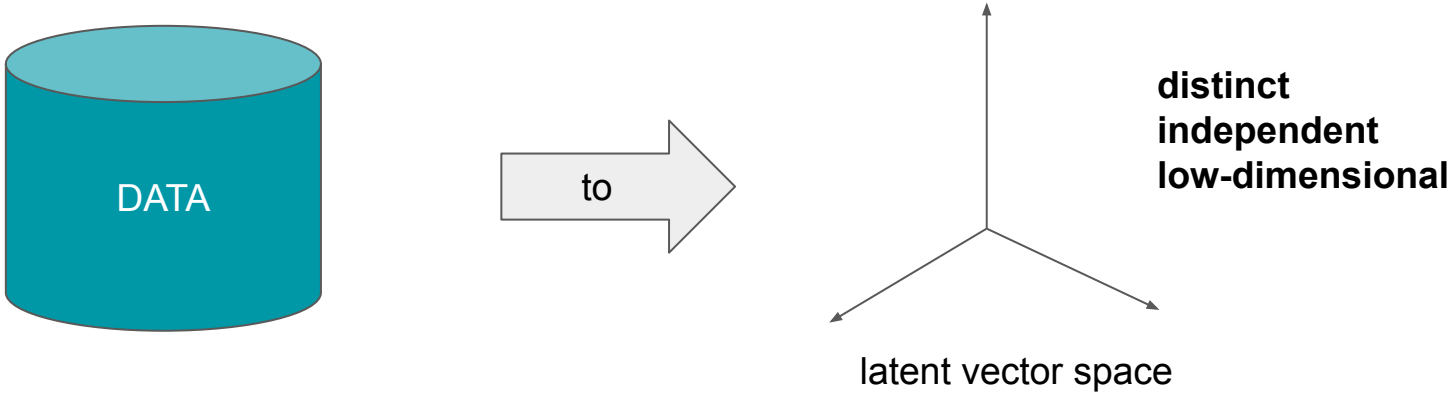
Lecture 9. Disentangled representation learning

InfoGAN · β-VAE · IDEL

Tatiana Makhalova
MADE, 13.11.2021

Disentangled representation learning



What does it mean to have distinct latent vectors?

Information theory. Mutual information

In disentangled representation learning, a common goal is to minimize the **mutual information** between different types of embeddings

Given two random variables $\boldsymbol{x}$ and $\boldsymbol{y}$, their MI is defined as

$$\mathrm{I}(\boldsymbol{x};\boldsymbol{y}) = \mathbb{E}_{p(\boldsymbol{x},\boldsymbol{y})}[\log \frac{p(\boldsymbol{x},\boldsymbol{y})}{p(\boldsymbol{x})p(\boldsymbol{y})}],$$

where $p(\boldsymbol{x}, \boldsymbol{y})$ is the joint distribution of the random variables, with $p(\boldsymbol{x})$and $p(\boldsymbol{y})$ representing the respective marginal distributions

MI is a measure of "**dependence**" between two variables

Drawback: expectation is usually intractable in practice

Information theory. Variation of Information

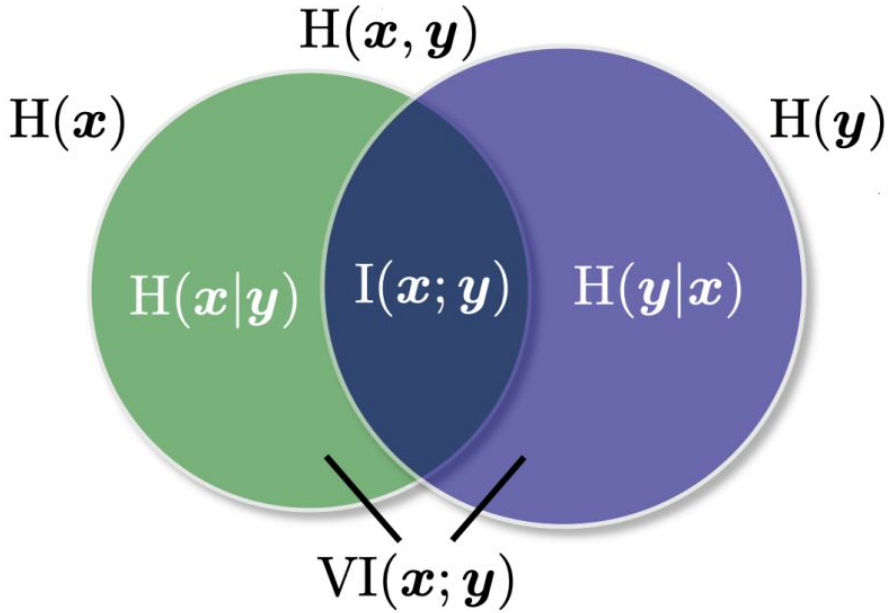
Variation of Information VI (or **Shared Information Distance**) is a measure of **independence** between two random variables

For random variables x and y VI is given as $\mathrm{VI}(x; y) = \mathbf{H}(x) + \mathbf{H}(y) - 2I(x; y)$
where $\mathbf{H}(x)$ and $\mathbf{H}(y)$ are entropies of x and y, respectively

Entropy is given by $\mathrm{H}(\boldsymbol{x}) = \mathbb{E}_{p(\boldsymbol{x})}[-\log p(\boldsymbol{x})]$

VI is a well-defined metric (identity of indiscernibles, symmetry, triangle inequality)

Different measures of information



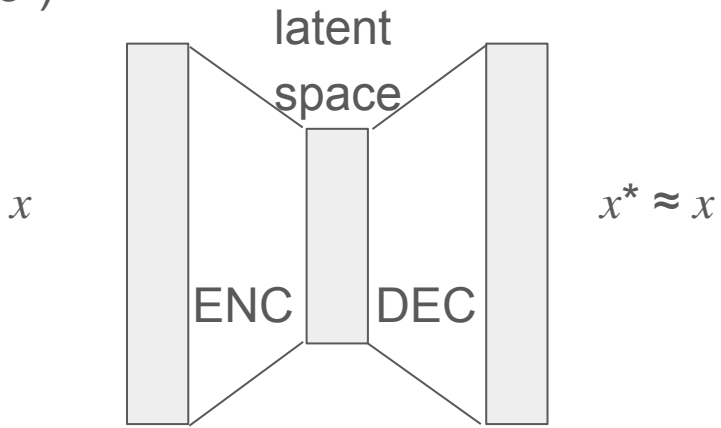
$$I(X;Y) = H(X) - H(X|Y) = H(Y) - H(Y|X)$$

Autoencoders

An **autoencoder** is a type of artificial neural network used to learn efficient codings of unlabeled data (unsupervised learning)

What AE **does:** regenerates the input

What AE **learns**: a representation (encoding) by training the network to ignore insignificant data (or “noise”)



Generative models

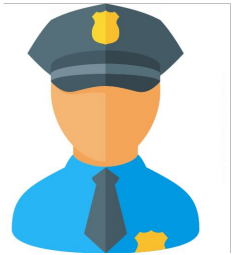
Generative Adversarial Networks is a game-theoretic approach. It can be tricky and unstable to tran, no inference queries

Variational Autoencoders (VAE) optimizes the variational lower bound on likelihood. Useful latent representation, inference queries. Sample quality might not be the best

Generative Adversarial networks. Basic idea

Two-player game:

G tries to fool D

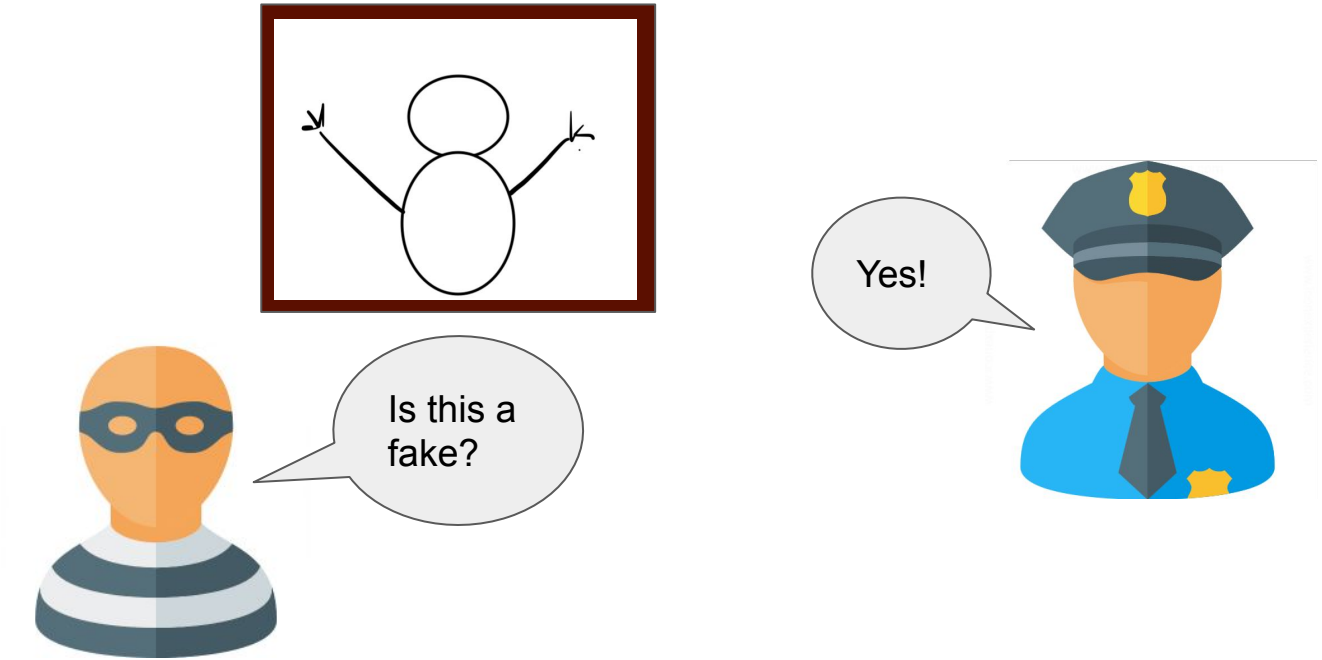


D tries to differentiate the true and fake instances

G is learning how to fool D,
e.g., how to draw this picture

Generative Adversarial networks. Basic idea

Beginning of learning



Generative Adversarial networks. Basic idea

In the middle of learning

Generative Adversarial networks. Basic idea

In the middle of learning

Objective

Train jointly in **minimax game**

$$\min_{\theta_g} \max_{\theta_d} [\mathbb{E}_{x \sim p_{data}} \log D_{\theta_d}(x) + \mathbb{E}_{z \sim p(z)} \log(1 - D_{\theta_d}(G_{\theta_g}(x)))]$$

$D_{\theta_d}(x)$: discriminator output for real data x

$D_{\theta_d}(G_{\theta_g}(x))$: discriminator output for generated fake data $G(z)$

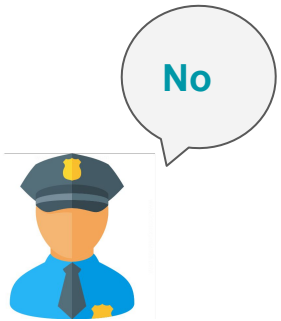
$$V(D_{\theta_d}, G_{\theta_g})$$

Goodfellow, Ian; Pouget-Abadie, Jean; Mirza, Mehdi; Xu, Bing; Warde-Farley, David; Ozair, Sherjil; Courville, Aaron; Bengio, Yoshua (2014). Generative Adversarial Nets. Proceedings of the International Conference on Neural Information Processing Systems (NIPS 2014). pp. 2672–2680

The goal of the Discriminator

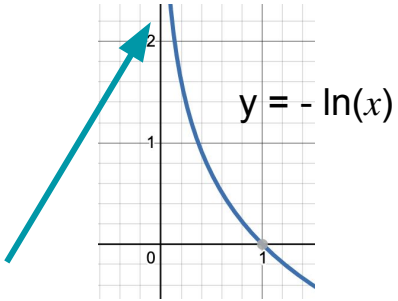
$$\min_{\theta_g} \max_{\theta_d} [\mathbb{E}_{x \sim p_{data}} \log D_{\theta_d}(x) + \mathbb{E}_{z \sim p(z)} \log(1 - D_{\theta_d}(G_{\theta_g}(x)))]$$

Is $x \sim p_{data}$ fake?

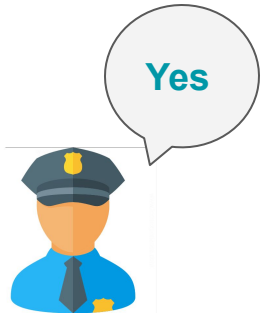


$$\log D_{\theta_d}(x)$$

0

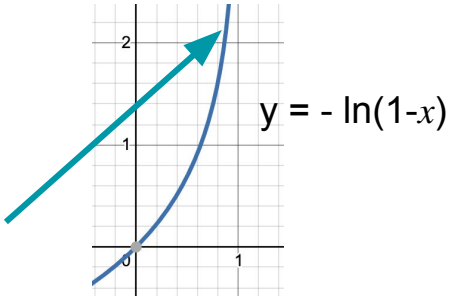


Is $x \sim p_z$ fake?



$$\log(1 - D_{\theta_d}(G_{\theta_g}(x)))$$

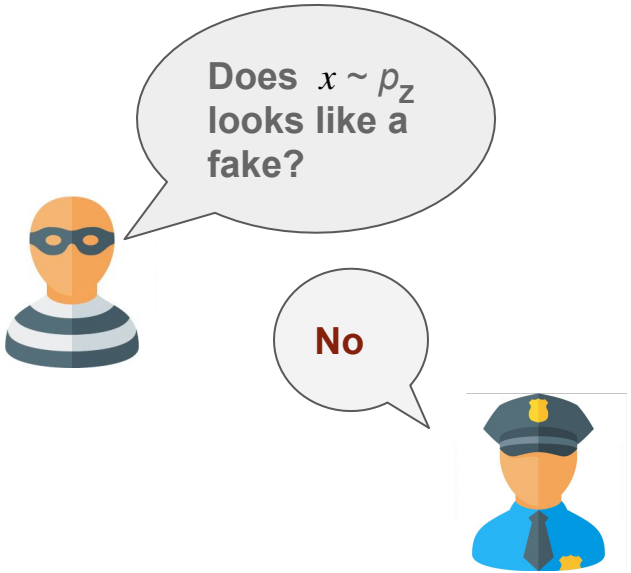
1



The goal of the Generator

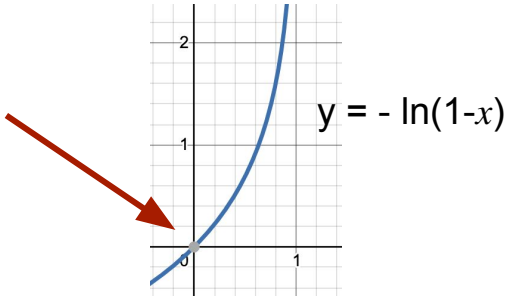
V(D, G)

$$\min_{\theta_g} \max_{\theta_d} \left[\mathbb{E}_{x \sim p_{data}} \log D_{\theta_d}(x) + \mathbb{E}_{z \sim p(z)} \log(1 - D_{\theta_d}(G_{\theta_g}(x))) \right]$$



$$\log(1 - D_{\theta_d}(G_{\theta_g}(x)))$$

1



Algorithm

for number of training iterations **do**
    for k steps **do**
        sample minibatch of m noise samples from $p_g(z)$
        sample minibatch of m examples from generating data distribution $p_{\text{data}}(z)$
        update the discrimination by ascending its stohastic gradient:

$$\nabla_{\theta_d} \frac{1}{m} \sum_{i=1}^{m} [\log D_{\theta_d}(x^{(i)}) + \log(1 - D_{\theta_d}(G_{\theta_g}(z^{(i)})))]$$

    end for
simple minimatch of noise samples from $p_g(z)$
update the generator by ascending its stochastic gradient

$$\nabla_{\theta_d} \frac{1}{m} \sum_{i=1}^{m} \log(D_{\theta_d}(G_{\theta_g}(z^{(i)})))$$

end for

InfoGAN

Idea : rather than using a single unstructured noise vector in G, it is decomposed into two components:

- z, which is treated as source of incompressible noise;
- c, which we will call the latent code and will target the salient structured semantic features of the data distribution.

Mathematically, we denote the set of structured latent variables by $c_1, c_2, \ldots, c_L$

We may assume a factored distribution, given by $P(c_1, c_2, \ldots, c_L) = \prod_{i=1}^{L} P(c_i)$

For ease of notation, further we use $c = (c_1, c_2, \ldots, c_L)$

Chen, X., Duan, Y., Houthooft, R., Schulman, J., Sutskever, I., & Abbeel, P. (2016, December). Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In *Proceedings of the 30th International Conference on Neural Information Processing Systems* (pp. 2180-2188)

InfoGAN

The idea is to make latent representation **diverse**. It could be done by minimising the mutual information between latent vectors:

$$I(X;Y) = H(X) - H(X|Y) = H(Y) - H(Y|X)$$

Thus, the modified objective is given by

$$\min_G \max_D V_I(D,G) = V(D,G) - \lambda I(c; G(z,c))$$

where c is a set of latent variables, and z is a source of incompressible noise

InfoGAN

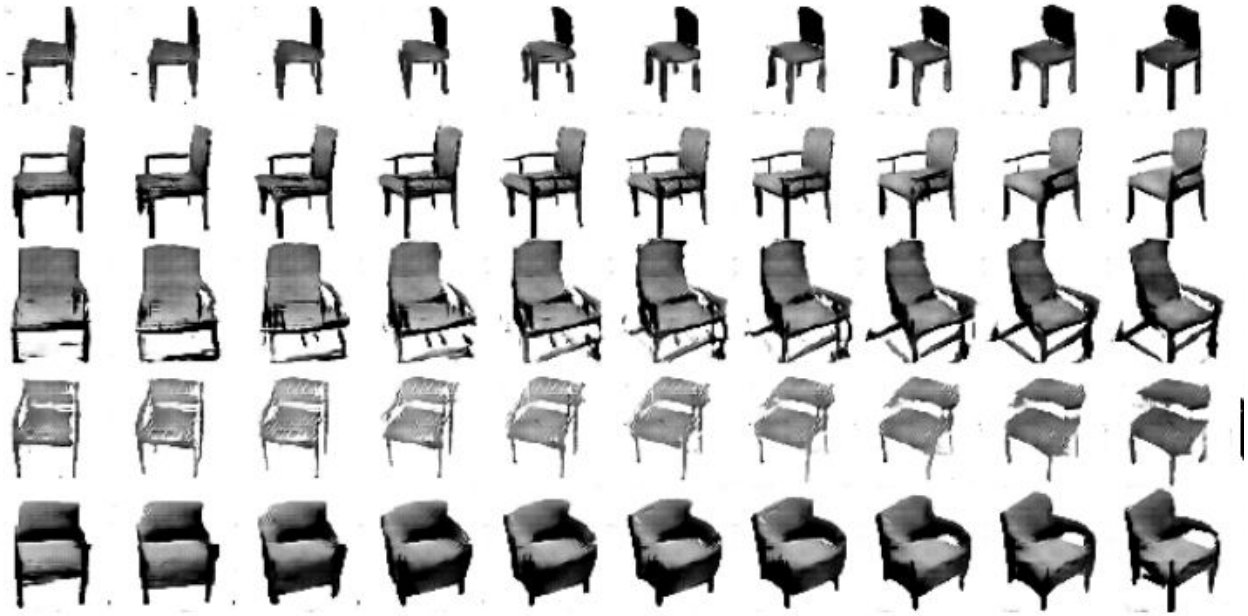
In practice $I(c,G(z,\ c))$ is intractable, thus a lower bound of MI is hard to maximize directly as it requires access to the posterior $P\ (c|x)$

Fortunately we can obtain a lower bound of it by defining an auxiliary distribution Q $(c|x)$ to approximate $P(c|x)$:

$$\begin{aligned} L_I(G,Q) &= E_{c\sim P(c),x\sim G(z,c)}[\log Q(c|x)] + H(c) \\ &= E_{x\sim G(z,c)}[\mathbb{E}_{c'\sim P(c|x)}[\log Q(c'|x)]] + H(c) \\ &\leq I(c;G(z,c)) \end{aligned}$$

Thus, the objective is given by $\min_{G,Q}\max_{D} V_{\text{InfoGAN}}(D,G,Q) = V(D,G) - \lambda L_I(G,Q)$

Example. Manipulating latent codes on 3D Chairs



(a) Rotation

Example. Manipulating latent codes on 3D Chairs

(b) Width

Manipulating latent codes on CelebA

(a) Azimuth (pose)

Manipulating latent codes on CelebA

(b) Presence or absence of glasses

Manipulating latent codes on CelebA

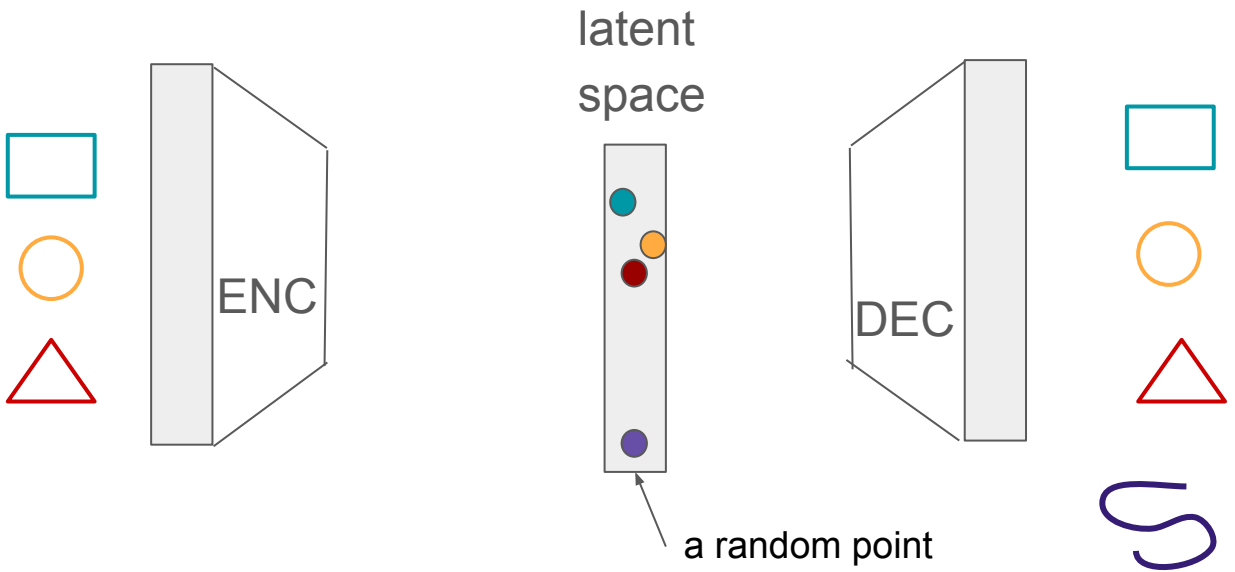
(c) Hair style

Manipulating latent codes on CelebA

(d) Emotion

GAN issues

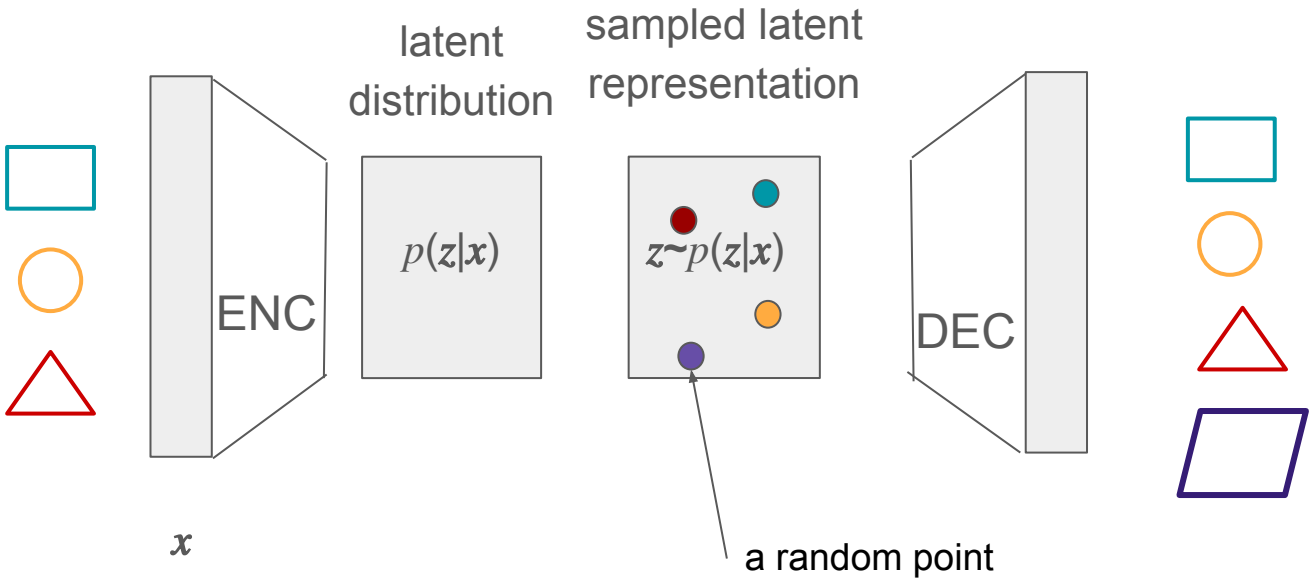
With GANs it is almost impossible to use only **deconder**, since it is unclear how the latent space is organized



How to enable the generation process?

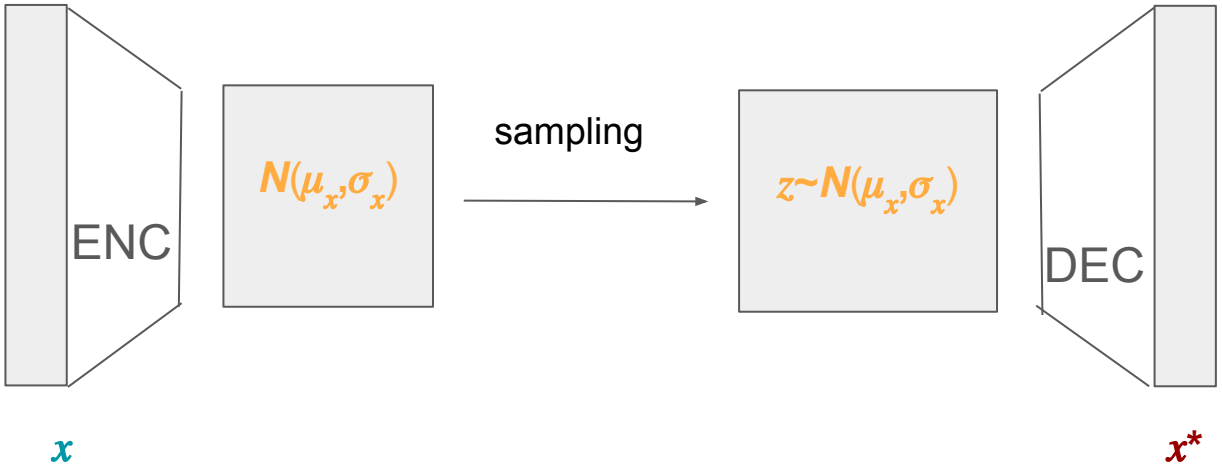
Variational autoencoders

With GANs it is almost impossible to use only deconder, since it is unclear how the latent space is organized



How to enable the generation process?

Variational autoencoders



$$\text{loss} = \| x - x^* \| + KL(N(\mu_x,\sigma_x), N(0,1)) = \| x - d(z) \| + KL(N(\mu_x,\sigma_x), N(0,1))$$

Kingma, Diederik P.; Welling, Max (2014-05-01). "Auto-Encoding Variational Bayes". arXiv:1312.6114

More about KL penalty term

Why do we need the KL divergence w.r.t. the standard normal distribution?

Since we want the latent space be regular

The **regular latent space** is the space that satisfies the following properties:

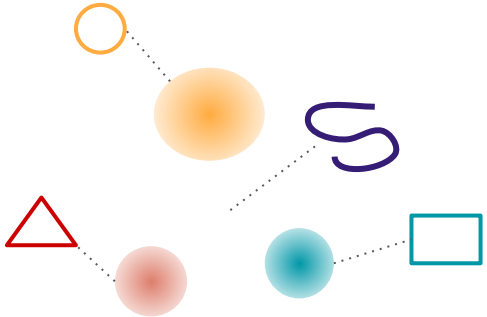
- **continuity** (two close points in the latent space should not give two completely different contents once decoded)
- **completeness** (for a chosen distribution, a point sampled from the latent space should give “meaningful” content once decoded)

More about KL penalty term

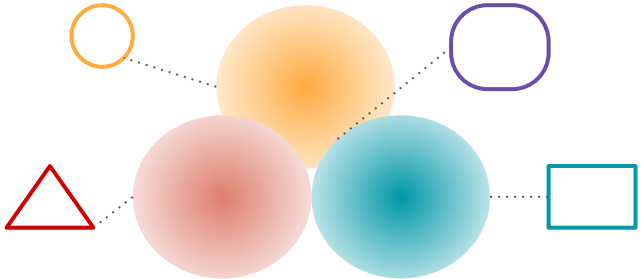
$$\text{loss} = \| x - x^* \| + \text{KL}(N(\mu_x, \sigma_x), N(0,1))$$

The encoder may return

- distributions with tiny variances (that would tend to be punctual distributions)
- distributions with very different means (that would then be really far apart from each other in the latent space)



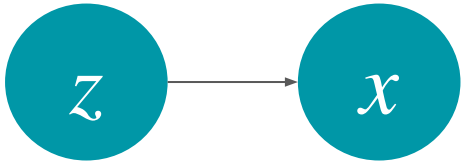
without penalty



with penalty

VAE more formally

Suppose that each data instance x is generated as follows:



latent representation is sampled from $p(z)$

data is sampled from the conditional likelihood distribution $p(x|z)$

Assuming that

$$p(z) \equiv \mathcal{N}(0, I)$$
$$p(x|z) \equiv \mathcal{N}(f(z), cI) \qquad f \in F \qquad c > 0$$

whose mean is defined by a deterministic function f of the variable of z and whose covariance matrix has the form of a positive constant c that multiplies the identity matrix I

VAE more formally

Using the Bayes theorem that makes the link between the prior $p(z)$, the likelihood $p(x|z)$, and the posterior $p(z|x)$, we get

$$p(z|x) = \frac{p(x|z)p(z)}{p(x)} = \frac{p(x|z)p(z)}{\int p(x|u)p(u)du}$$

However, $p(z|x)$ can not be found using a classical Bayesian inference problem, since computing the integral is intractable in practice

How to compute $p(z|x)$ then?

Variational inference

VI is a technique to approximate complex distributions

Idea: is to set a parameterized family of distribution and to look for the best approximation of our target distribution among this family

Here $p(x|z)$ is approximated by a Gaussian distribution $q_x(z)$ whose mean and covariance are defined by two functions, g and h, of the parameter x

$$q_x(z) \equiv \mathcal{N}(g(x), h(x)) \qquad g \in G \qquad h \in H$$

Variational inference

To find the best approximation among this family, we do the following:

$$
\begin{aligned}
(g^*, h^*) &= \underset{(g,h)\in G\times H}{\arg\min}\ KL(q_x(z), p(z|x)) \\
&= \underset{(g,h)\in G\times H}{\arg\min}\ \left(\mathbb{E}_{z\sim q_x}(\log q_x(z)) - \mathbb{E}_{z\sim q_x}\left(\log \frac{p(x|z)p(z)}{p(x)}\right)\right) \\
&= \underset{(g,h)\in G\times H}{\arg\min}\ (\mathbb{E}_{z\sim q_x}(\log q_x(z)) - \mathbb{E}_{z\sim q_x}(\log p(z)) - \mathbb{E}_{z\sim q_x}(\log p(x|z)) + \mathbb{E}_{z\sim q_x}(\log p(x))) \\
&= \underset{(g,h)\in G\times H}{\arg\max}\ (\mathbb{E}_{z\sim q_x}(\log p(x|z)) - KL(q_x(z), p(z))) \\
&= \underset{(g,h)\in G\times H}{\arg\max}\ \left(\mathbb{E}_{z\sim q_x}\left(-\frac{||x - f(z)||^2}{2c}\right) - KL(q_x(z), p(z))\right)
\end{aligned}
$$

maximising the likelihood of the "observations"

staying close to the prior distribution

Variational inference

In practice the function f, that defines the decoder, is not known and also need to be chosen

In fact, for a given input x, we want to maximise the probability to have $x^* = x$ when we sample z from the distribution $q^*_x(z)$ and then sample x^* from the distribution $p(x|z)$

$$f^* = \arg\max_{f \in F} \mathbb{E}_{z \sim q^*_x} (\log p(x|z))$$

$$= \arg\max_{f \in F} \mathbb{E}_{z \sim q^*_x} \left(-\frac{||x - f(z)||^2}{2c} \right)$$

VAE and β-VAE

Gathering all the pieces together, the objective for VAE is the following

$$(f^*, g^*, h^*) = \underset{(f,g,h) \in F \times G \times H}{\arg\max} \left(\mathbb{E}_{z \sim q_x} \left(-\frac{||x - f(z)||^2}{2c} \right) - KL(q_x(z), p(z)) \right)$$

To make the latent representation more diverse, we put a greater penalty on KL divergence

$$(f^*, g^*, h^*) = \underset{(f,g,h) \in F \times G \times H}{\arg\max} \left(\mathbb{E}_{z \sim q_x} \left(-\frac{||x - f(z)||^2}{2c} \right) - \beta KL(q_x(z), p(z)) \right)$$

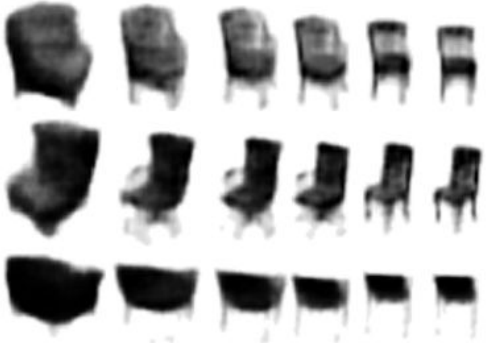
Example. 3D Chairs. “Azimuth” variable

β = 5

Example. 3D Chairs. “Width” variable

InfoGAN

β-VAE



VAE

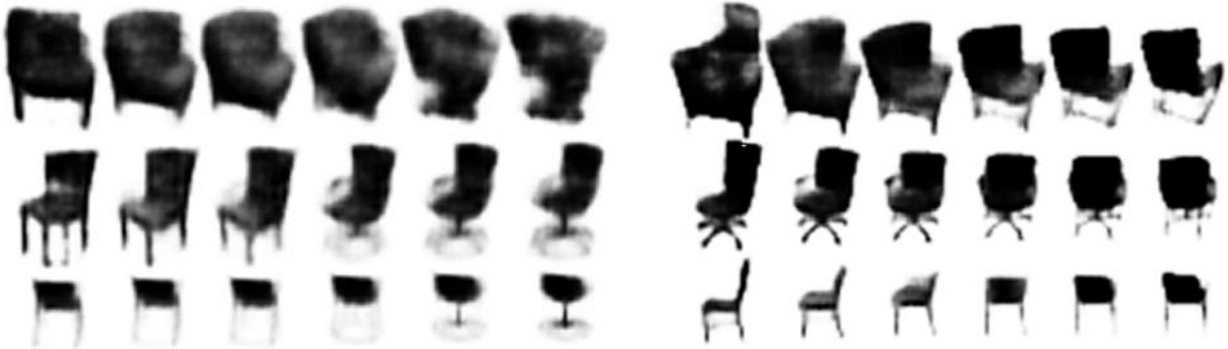
Example. 3D Chairs. “Leg style” variable

InfoGAN

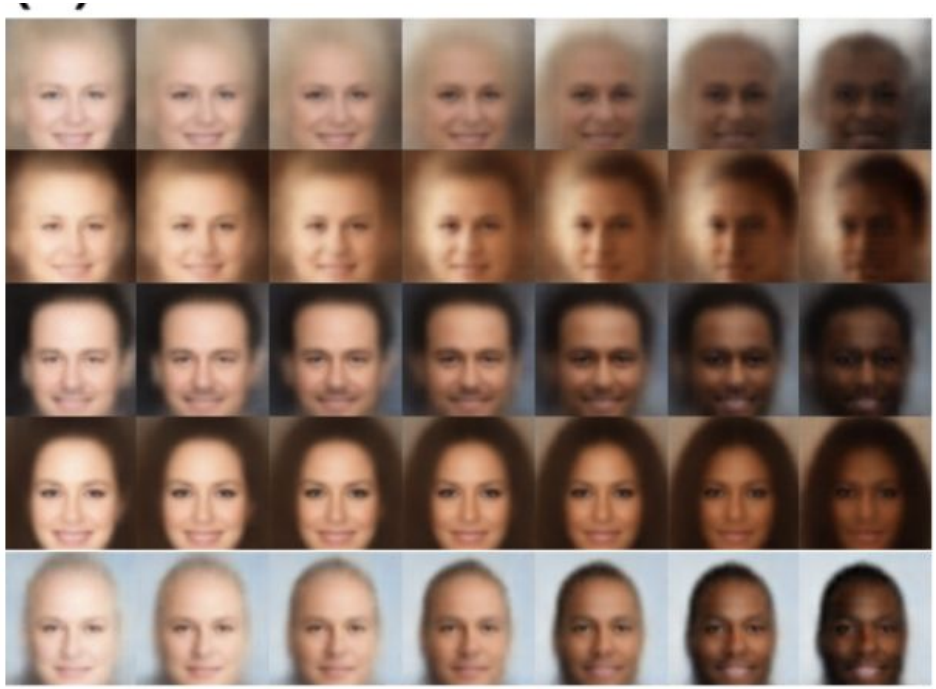
β-VAE

VAE

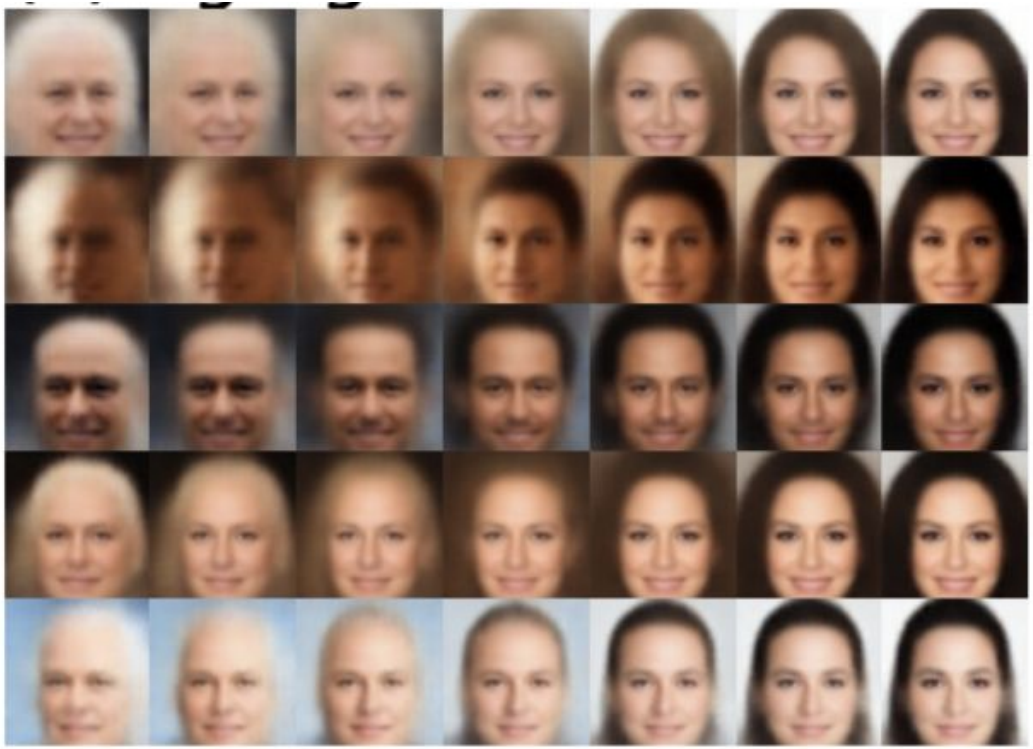
Factor not learnt



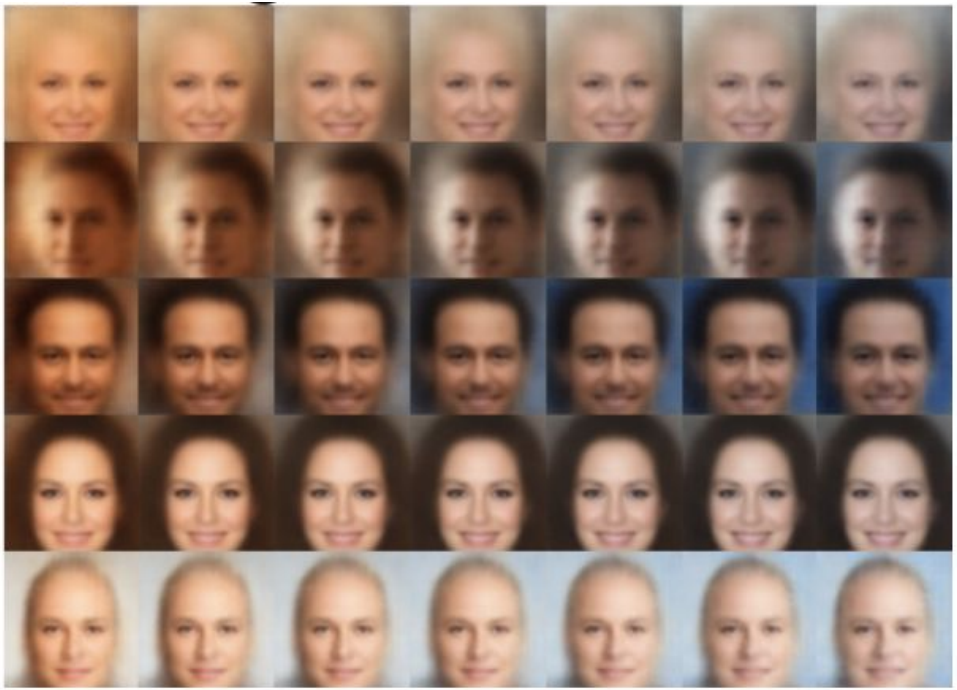
Example of β-VAE. CelebA. Skin color



Example of β-VAE. CelebA. Age / gender



Example of β-VAE. CelebA. Image saturation



Natural Language Processing

- style embedding
  - represent desired attributes, such as the sentiment of a review, or the personality associated with a post
- content embedding
  - designed to encapsulate the semantic meaning of a sentence. In contrast, the style embedding should

Ideally, a disentangled-text-representation model should learn representative embeddings for both style and content

IDEL. Objective

The goal is to encode each sentence x_i into its corresponding style embedding s_i and content embedding c_i with an encoder $q_\theta(s,c|x)$ such that $s_i,c_i|x_i \sim q_\theta(s,c|x)$

Our overall disentangled representation learning objective is:

$$\min L_{\text{Dis}} = I(s,c) - I(c,x) - I(s,y)$$

maximize to ensure that the content embedding c sufficiently encapsulates information from the sentence x

the same reasoning

Theoretical justification of the measure

The objective L_{Dis} has a strong connection with VI(x; y) = **H**(x) + **H**(y) − 2I(x; y)

Applying the triangle inequality to s, c and x, we have VI(s; x) + VI(x; c) ≥ VI(s; c)

the degree of disentanglement is represented as follows:s,

$$D(\boldsymbol{x};\boldsymbol{s},\boldsymbol{c}) = \mathrm{VI}(\boldsymbol{s};\boldsymbol{x}) + \mathrm{VI}(\boldsymbol{x};\boldsymbol{c}) - \mathrm{VI}(\boldsymbol{c};\boldsymbol{s}) = 2\mathrm{H}(\boldsymbol{x}) + 2[\mathrm{I}(\boldsymbol{s};\boldsymbol{c}) - \mathrm{I}(\boldsymbol{x};\boldsymbol{c}) - \mathrm{I}(\boldsymbol{x};\boldsymbol{s})]$$

a constant associated with the data, can be removed

D(x; s, c) is symmetric to style s and content c, thus it can be difficult to separate and distinguish them

Theoretical justification of the measure

Let us consider the following dependence $s \rightarrow x \rightarrow y$ (it is a Markov Chain), thus $I(s;x) \geq I(s; y)$[1].

Then $I(s;c) - I(x;c) - I(x;s) \geq I(s;c) - I(x;c) - I(y; s)$ which gives $\boldsymbol{L}_{\text{Dis}}$

Minimizing the exact value of mutual information in $\boldsymbol{L}_{\text{Dis}}$ causes **numerical instabilities**, especially when the dimension of the latent embeddings is large

1. based on the MI data-processing inequality, Thomas M Cover and Joy A Thomas. 2012. Elements of information theory. John Wiley & Sons
2. Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. 2016. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In Advances in neural information processing systems, pages 2172–2180.

Variational lower bound

For I(x;c), we introduce a variational decoder $q_\varphi(x|c)$ to reconstruct the sentence x by the content embedding c

With a variational distribution we can derive $\mathbf{I}(\boldsymbol{x};\boldsymbol{y}) \geq \mathbf{H}(\boldsymbol{x}) + \mathbb{E}_{p(\boldsymbol{x},\boldsymbol{y})}[\log q(\boldsymbol{x}|\boldsymbol{y})]$

Using the formula above, I(x;c) ≥ H(x) + E_p(x;c) [log $q_\varphi(x|c)$]

Similarly, for I(s; y): I(s;y) ≥ H(y) + E_p(y,s) [log $q_\psi(y|s)$], where $q_\psi(y|s)$ is a classifier mapping the style embedding s to its corresponding style label y

Thus, the lower bound for $\boldsymbol{L}_{\text{Dis}}$ is given by

$$\mathcal{L}_{\text{Dis}} \leq \mathbf{I}(\boldsymbol{s};\boldsymbol{c}) - [\mathbf{H}(\boldsymbol{x}) + \mathbb{E}_{p(\boldsymbol{x},\boldsymbol{c})}[\log q_\phi(\boldsymbol{x}|\boldsymbol{c})]] \\ -[\mathbf{H}(y) + \mathbb{E}_{p(y,\boldsymbol{s})}[\log q_\psi(y|\boldsymbol{s})]]$$

Variational lower bound

Thus, we need to minimize

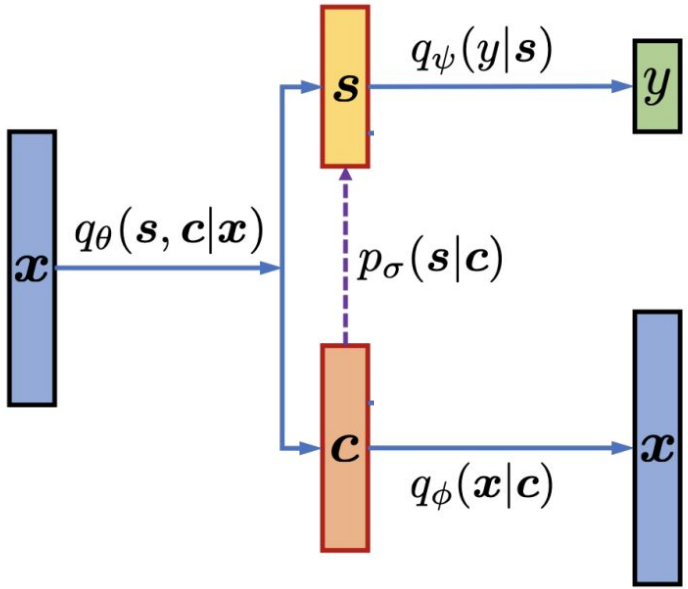
$$\bar{\mathcal{L}}_{\text{Dis}} = \mathrm{I}(\boldsymbol{s};\boldsymbol{c}) - \mathbb{E}_{p(\boldsymbol{x},\boldsymbol{c})}[\log q_\phi(\boldsymbol{x}|\boldsymbol{c})] - \mathbb{E}_{p(y,\boldsymbol{s})}[\log q_\psi(y|\boldsymbol{s})]$$

Where I(s,c) can be estimated as well (a lot of theory)

The architecture of the solution

Each sentence x is encoded into style embedding s and content embedding c

The style embedding s goes through a classifier $q_\psi(y|s)$ to predict the style label y; the content embedding c is used to reconstruct x

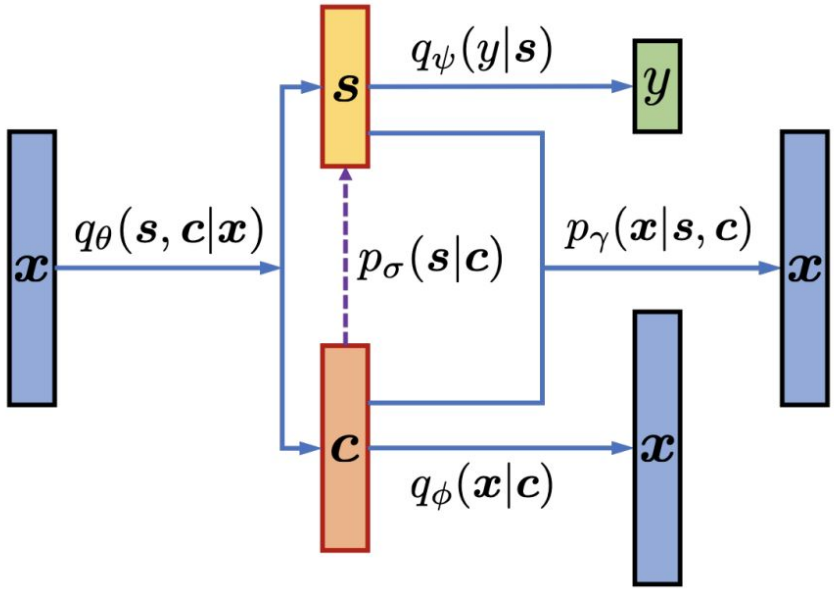


Building a VAE

Since the proposed DRL encoder $q_{\varphi}(x|c)$ is a stochastic neural network, a natural extension is to add a decoder to build a variational autoencoder (VAE)

Let $p_{\gamma}(x|s,c)$ be **a decoder network** that generates a new sentence based on the given style s and content c

The decoder $p_{\gamma}(x|s,c)$ generates sentences based on the combination of s and c



Building a VAE

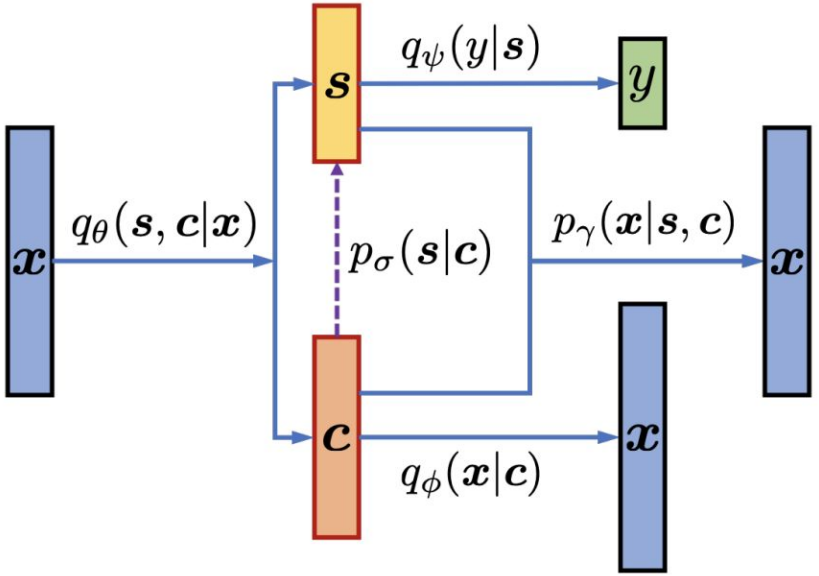
A prior distribution p(s,c) = p(s)p(c), as the product of two multivariate unit-variance Gaussians, is used to regularize the posterior distribution $q_\theta(s,c|x)$ by KL-divergence minimization

Meanwhile, the log-likelihood term for text reconstruction should be maximized. The objective for VAE is:

$$\begin{aligned}\mathcal{L}_{\text{VAE}} =& \mathbf{KL}(q_\theta(\boldsymbol{s}, \boldsymbol{c}|\boldsymbol{x}) \| p(\boldsymbol{s}, \boldsymbol{c})) \\ & - \mathbb{E}_{q_\theta(\boldsymbol{s},\boldsymbol{c}|\boldsymbol{x})}[\log p_\gamma(\boldsymbol{x}|\boldsymbol{s}, \boldsymbol{c})]\end{aligned}$$

Information-theoretic Disentangled text Embedding Learning

The final loss function is $L_{total} = \beta L^{*}_{Dis} + L_{VAE}$



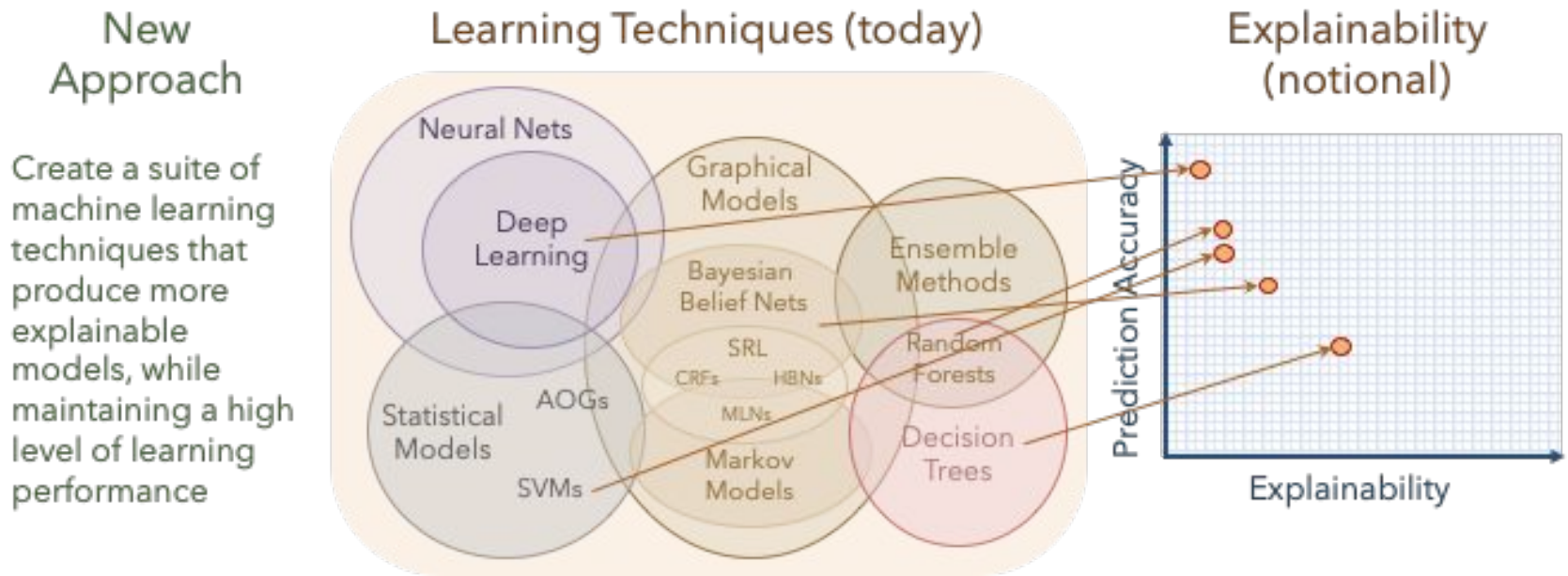
More examples of DRL

- domain adaptation
  - Yen-Cheng Liu, Yu-Ying Yeh, Tzu-Chien Fu, Sheng-De Wang, Wei-Chen Chiu, and Yu-Chiang Frank Wang. 2018. Detach and adapt: Learning cross-domain disentangled deep representation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 8867–8876
- style transfer
  - Hsin-Ying Lee, Hung-Yu Tseng, Jia-Bin Huang, Maneesh Singh, and Ming-Hsuan Yang. 2018. Diverse image-to-image translation via disentangled representations. In Proceedings of the European Conference on Computer Vision (ECCV), pages 35–51
- conditional generation
  - Emily L Denton et al. 2017. Unsupervised learning of disentangled representations from video. In Advances in neural information processing systems, pages 4414–4423
  - Christopher P Burgess, Irina Higgins, Arka Pal, Loic Matthey, Nick Watters, Guillaume Desjardins, and Alexander Lerchner. 2018. Understanding disentangling in beta-vae. arXiv preprint arXiv:1804.03599
- images
  - Luan Tran, Xi Yin, and Xiaoming Liu. 2017. Disentangled representation learning gan for pose-invariant face recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 1415–1424
  - Hsin-Ying Lee, Hung-Yu Tseng, Jia-Bin Huang, Maneesh Singh, and Ming-Hsuan Yang. 2018. Diverse image-to-image translation via disentangled representations. In Proceedings of the European Conference on Computer Vision (ECCV), pages 35–51
- videos
  - Li Yingzhen and Stephan Mandt. 2018. Disentangled sequential autoencoder. In International Conference on Machine Learning, pages 5656–5665
  - Jun-Ting Hsieh, Bingbin Liu, De-An Huang, Li F FeiFei, and Juan Carlos Niebles. 2018. Learning to decompose and disentangle representations for video prediction. In Advances in Neural Information Processing Systems, pages 517–526
- speech
  - Juchieh Chou, Cheng chieh Yeh, Hung yi Lee, and Lin shan Lee. 2018. Multi-target voice conversion without parallel data by adversarially learning disentangled audio representations. In Proc. Interspeech 2018, pages 501–505
  - Hang Zhou, Yu Liu, Ziwei Liu, Ping Luo, and Xiaogang Wang. 2019. Talking face generation by adversarially disentangled audio-visual representation. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 33, pages 9299–9306

Final remarks

Accuracy vs Explainability

For a long time it is assumed that only unexplainable models may give good results



Credit: https://www.darpa.mil/attachments/XAIIndustryDay_Final.pptx, 2016

White-box models may be very accurate

The Rashomon effect occurs when many different explanations exist for the same phenomenon

In machine learning Leo Breiman coined this term to describe cases when there exist many different approximately-equally accurate models

$$rashomon\ ratio = \frac{|accurate\ models|}{|hypothesis\ space|}$$

Rashomon ratio provides a certificate of the existence of a simpler model that generalizes, rather than acting itself as a simplicity measure

Semenova, Lesia, Cynthia Rudin, and Ronald Parr. "A study in Rashomon curves and volumes: A new perspective on generalization and model simplicity in machine learning." *arXiv preprint arXiv:1908.01755* (2019)

Explainable accurate models

Even complex models can be explainable by its structure. In some cases, the objective may ensure interpretability:

Examples:

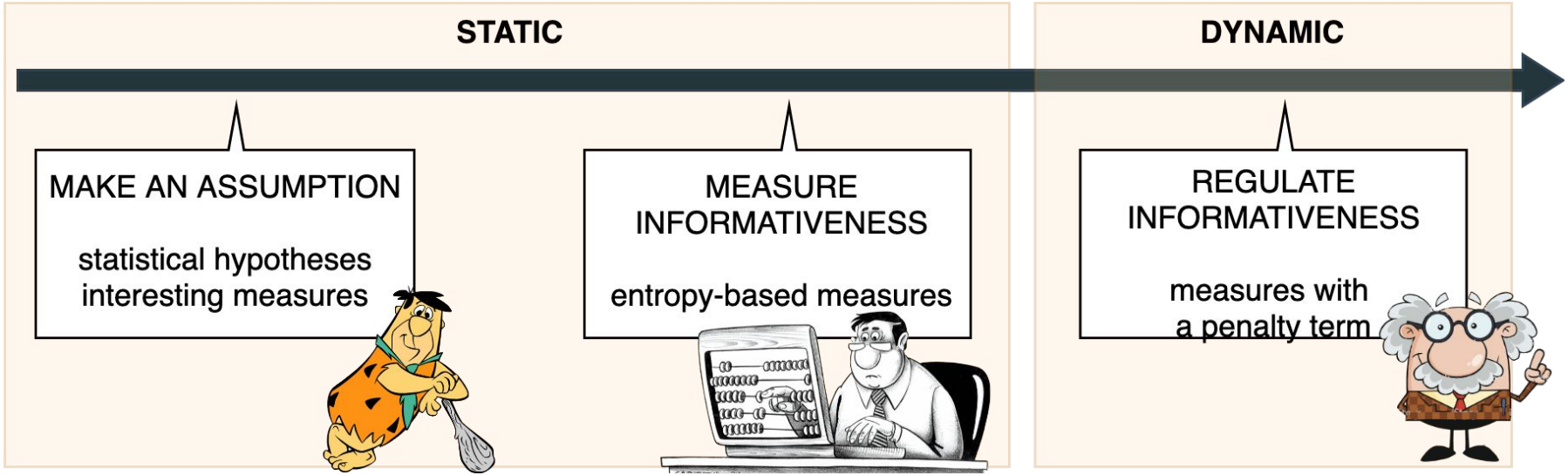
- Autoencoders: InfoGAN, VAE, IDEL
- Pattern mining methods: MDL-based approaches

In the majority of cases, understanding of the domain plays a crucial role

Examples:

- 2-layer additive risk model (an explainable analog of fully connected network)
- ProtoPNet (an explainable CNN)

Evolution of interestingness measures



Explainable accurate models

Even complex models can be explainable by its structure. In some cases, the objective may ensure interpretability:

Examples:

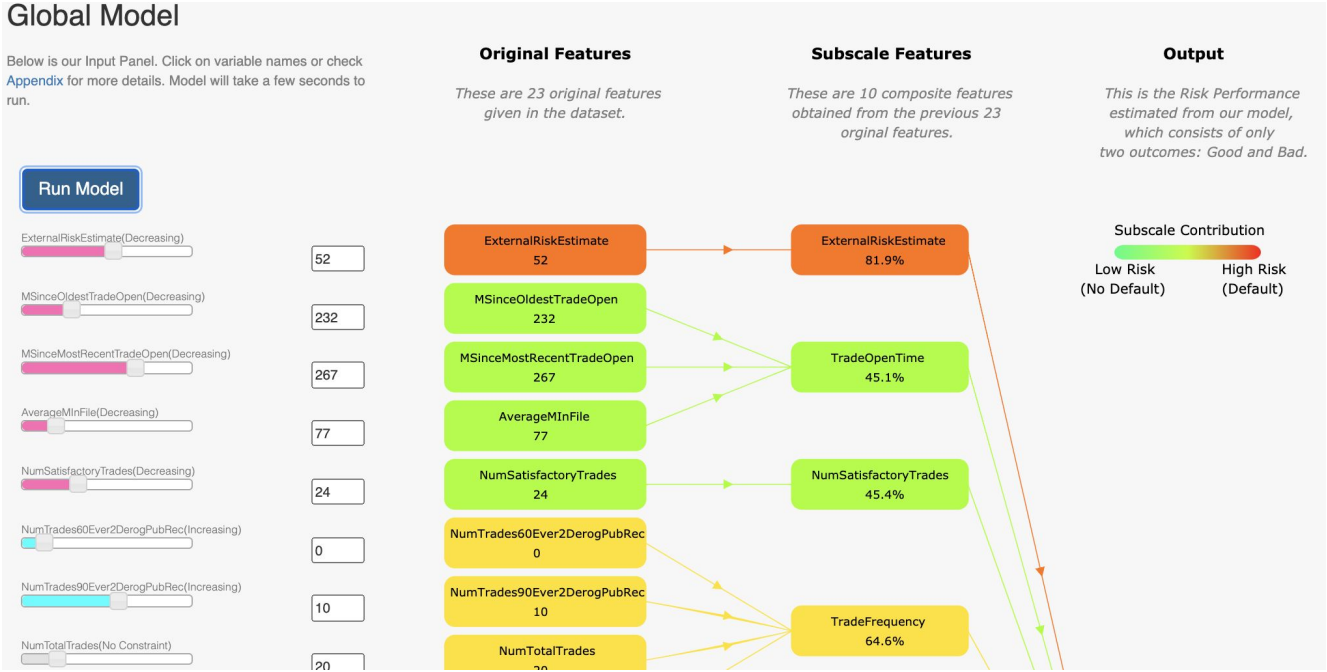
- Autoencoders: InfoGAN, VAE, IDEL
- Pattern mining methods: MDL-based approaches

In the majority of cases, understanding of the domain plays a crucial role

Examples:

- 2-layer additive risk model (an explainable analog of fully connected network)
- ProtoPNet (an explainable CNN)

2-layer additive risk model. An interpretable analog of NN



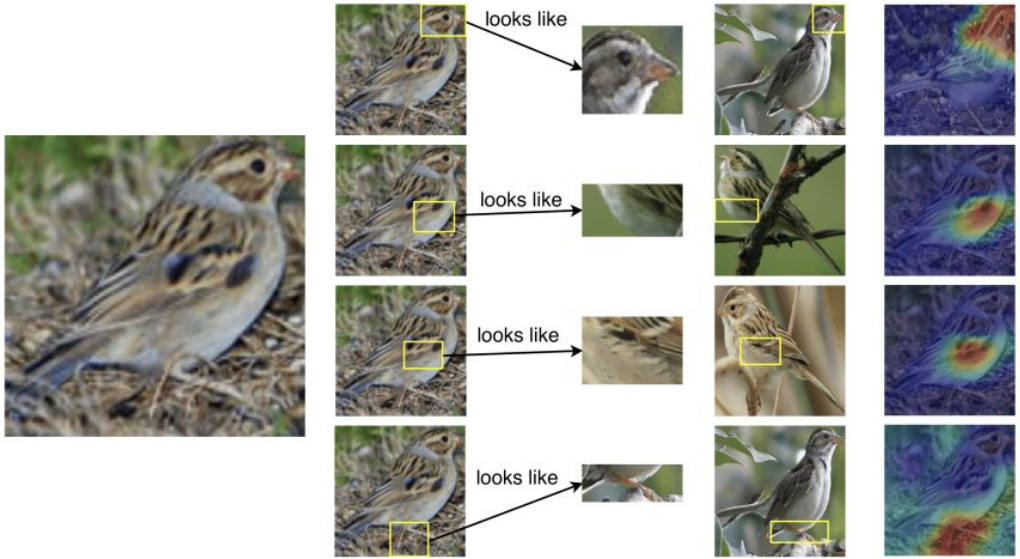
FICO explainable ML challenge for predicting credit risk

Detailed description: Chen C, Lin K, Rudin C, Shaposhnik Y, Wang S, Wang T. An Interpretable Model with Globally Consistent Explanations for Credit Risk. In: Proceedings of NeurIPS 2018 Workshop on Challenges and Opportunities for AI in Financial Services: the Impact of Fairness, Explainability, Accuracy, and Privacy; 2018

Source: http://dukedatasciencefico.cs.duke.edu/models/

ProtoPNet. An explainable CNN

Idea: imitate the reasoning process which is qualitatively similar to that of humans, namely paying attention to the part of images that are similar to the prototypes



Leftmost: a test image of a clay-colored sparrow
Second column: same test image, each with a bounding box generated by our model -- the content within the bounding box is considered by our model to look similar to the prototypical part (same row, third column) learned by our algorithm
Third column: prototypical parts learned by our algorithm
Fourth column: source images of the prototypical parts in the third column
Rightmost column: activation maps indicating how similar each prototypical part resembles part of the test bird

Chen, Chaofan, et al. "This looks like that: deep learning for interpretable image recognition." *arXiv preprint arXiv:1806.10574* (2018)

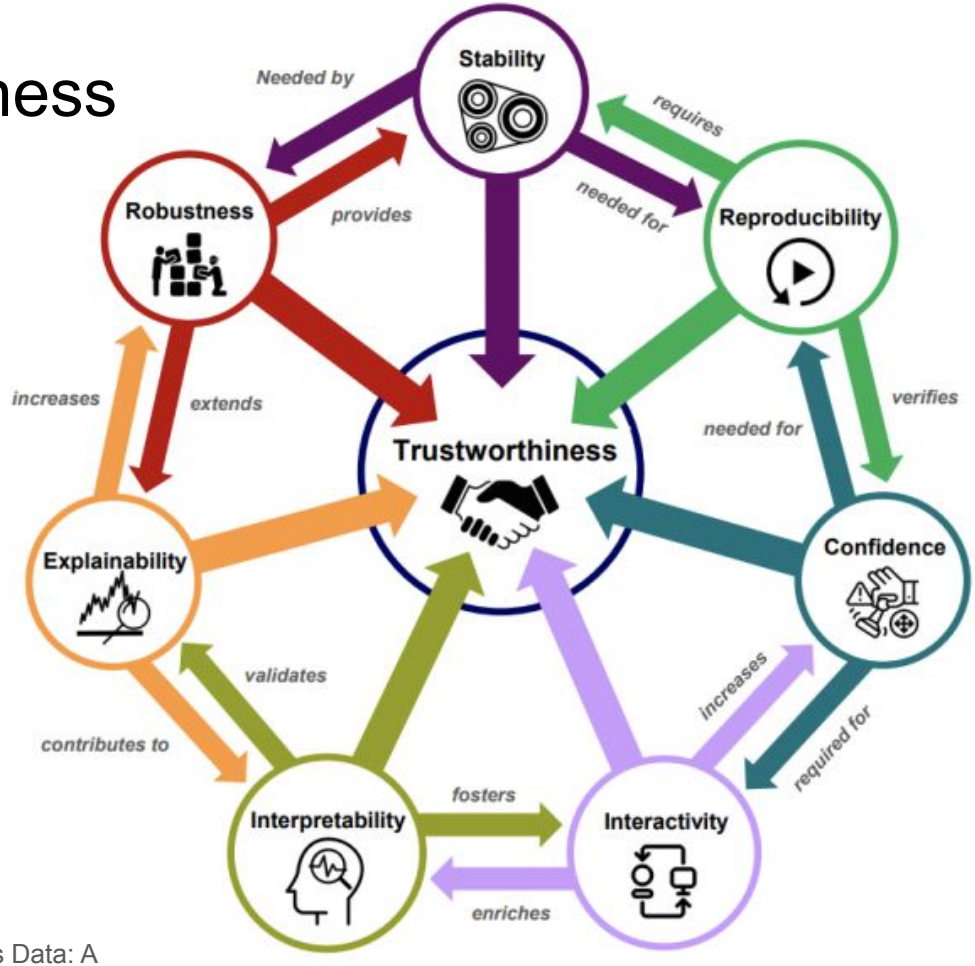
Pros and cons of interpretable models

+ “in-built” interpretability
+ trustable and transparent
+ easy to debug

- hard to find “good” features (often)
- requires domain understanding
- can be time-consuming (and iterative)

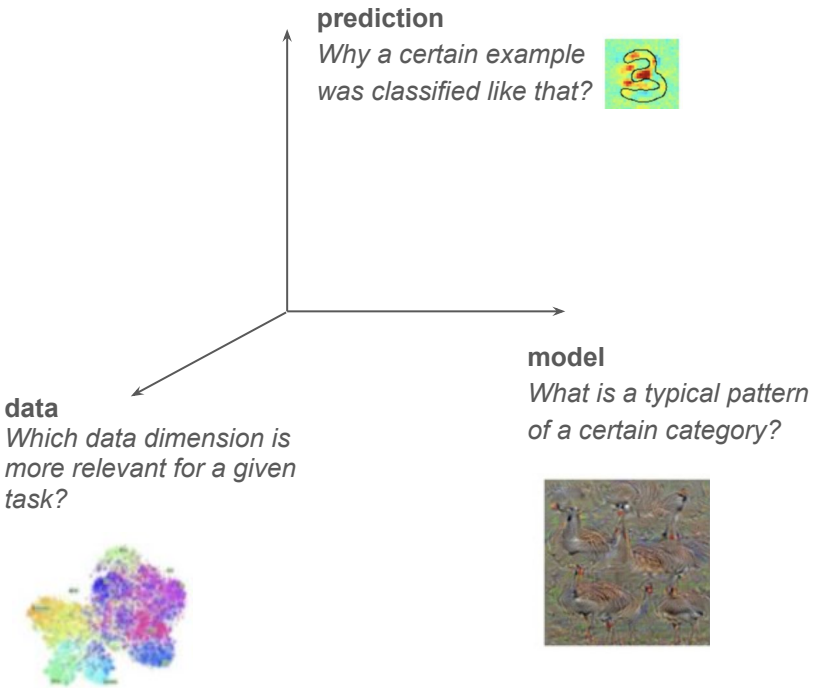
Components of trustworthiness

- Stability
- Robustness
- Reproducibility
- Confidence
- Interactivity
- Interpretability
- Explainability



Rojat, Thomas, et al. "Explainable Artificial Intelligence (XAI) on TimeSeries Data: A Survey." arXiv preprint arXiv:2104.00950 (2021)

Dimensions of explainability



MICCAI2018, tutorial Interpretable machine learning: http://iphome.hhi.de/samek/pdf/MICCAI2018.pdf

Explaining data

Model-independent explanation:

- Causal inference (“DoWhy” package)
  - correlation is not causality
  - the “treatment” effect can be measured when 2 groups are properly chosen

Model-related explanation:

- Dependencies between independent and dependent attributes (PDP, ALE, ICE plots)
- Feature importance and feature interaction (“SHAP method)

Explaining a particular prediction

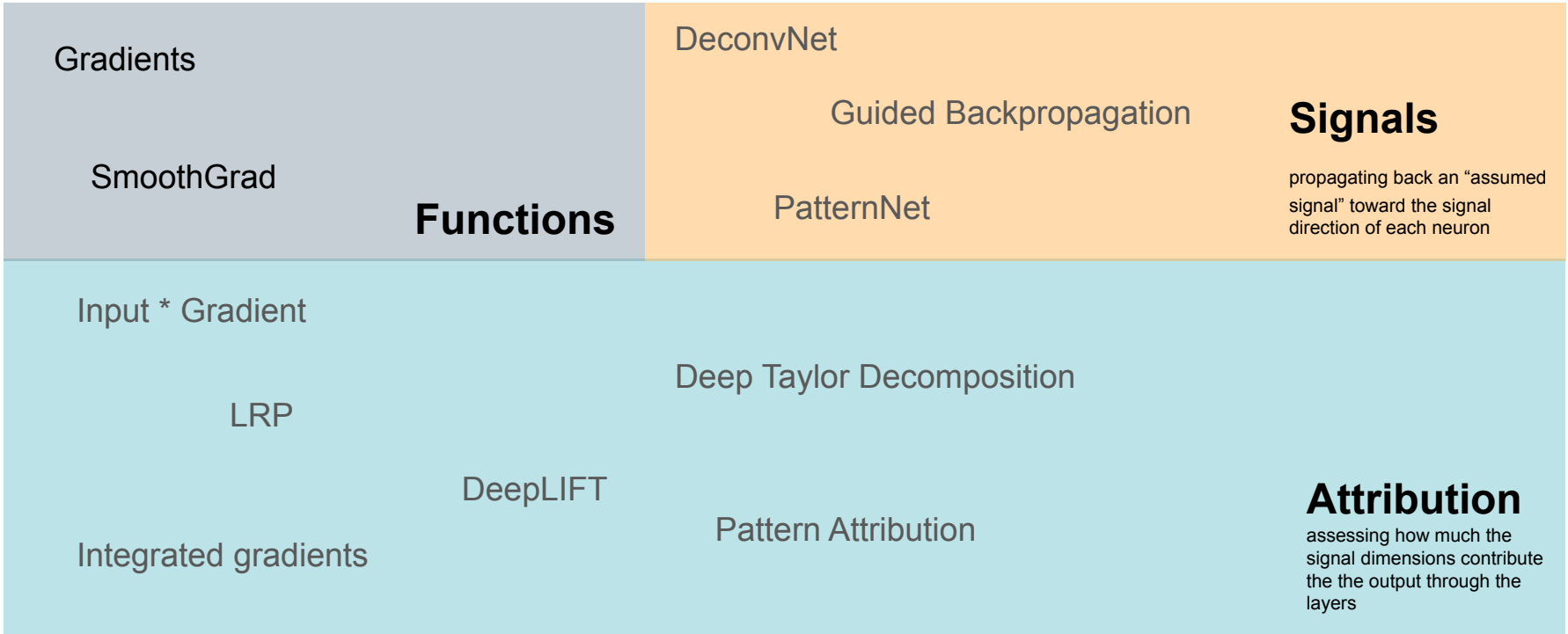
Model-agnostic approaches:

- methods based on counterfactuals
  - What are the minimal changes that are needed to be done to change the model prediction?
- LIME (local explanation method)
- SHAP (theory-based method)

Model-specific approaches:

- tree explainers (RuleFit, RuleScope, etc)
- neural network visualizers
- neural network explainers

Neural networks explainers



Explaining the model

- Activation maximization for neural networks
- Some visualizations provided by SHAP
- Submodular Pick in LIME

Pros and cons of explanation methods

+ universal (any models can be explained)
+ easy to train models
+ do not require any background knowledge

- may be contradictory
- unsupervised (no ground truth)
- lack of theoretical justification
- explain a model, but not the true relations between attributes
- explanation may be limited because of the simplicity of the explanation model

EU development concept for XAI

*"On many aspects however, AI systems that are currently under development are **far from achieving the minimal requirements of safety and security** that would be expected from autonomous systems."*

"As of now, several avenues for reflection could be considered to undertake the implementation of standards in AI technologies, and of security and reliability certifications of AI components embedded in real systems. These avenues include:

*— developing a **methodology to evaluate the impacts of AI systems on society** built on the model of the Data Protection Impact Assessments (DPIA) introduced in the GDPR, that would provide an assessment of the risks involved in the usage of AI models to the users and organisations;*

*— introducing **standardized tests to assess the robustness** of AI models, in particular to determine their field of action with respect to the data that have been used for the training, the type of mathematical model, and the context of use, amongst others factors;*

*— raising **awareness** among AI practitioners through the publication of **good practices regarding to known vulnerabilities** of AI models, and technical solutions to address them;*

*— promoting **transparency** in the conception of machine learning models, emphasizing the need of an **explainability-by-design** approach for AI systems with potential negative impacts on fundamental rights of users."*

Hamon, R., Junklewitz, H., Sanchez, I. Robustness and Explainability of Artificial Intelligence - From technical to policy solutions, EUR 30040, Publications Office of the European Union, Luxembourg, Luxembourg, 2020, ISBN 978-92-79-14660-5 (online), doi:10.2760/57493 (online), JRC119336.

EU rights for explanation

"*The data subject shall have the right* ***not to be subject to a decision based solely on automated processing****, including profiling, which produces legal effects concerning him or her or similarly significantly affects him or her*"

The latter is related to **the rights of the data subject** for

- understanding a certain decision
- contesting it
- altering to get the desired result

Article 22 of GDPR regulations [Onliine] https://www.privacy-regulation.eu/en/22.htm

Further reading

- MDL-based methods for pattern mining
- Causal inference (there is some stuff that we did not learn)
- Dealing with recurrent networks
  - **TimeSHAP** Bento, J., Saleiro, P., Cruz, A. F., Figueiredo, M. A., & Bizarro, P. (2021, August). TimeSHAP: Explaining recurrent models through sequence perturbations. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining* (pp. 2565-2573)
- XRL (Explanation in reinforcement learning)
  - Puiutta, E., & Veith, E. M. (2020, August). Explainable reinforcement learning: A survey. In International Cross-Domain Conference for Machine Learning and Knowledge Extraction (pp. 77-95). Springer, Cham
  - Heuillet, A., Couthouis, F., & Díaz-Rodríguez, N. (2021). Explainability in deep reinforcement learning. *Knowledge-Based Systems*, *214*, 106685

Stay tuned ! :)

Next lecture:

- The lecture starts at 10 a.m. the 27th of november
- Each presentation lasts 5 minutes (can be presented by any member of the team)
- https://docs.google.com/spreadsheets/d/1GAb36nNogTVQ1YVqsKolVz81UX3vBVmkkWVCaSldnpU/edit?usp=sharing