technische universität dortmund

Context: Undergraduate Project "Machine Learning" at TU Dortmund

Based on:

- "*Auto-Encoding Variational Bayes*" by Diederik P. Kingma and MaxWelling
- "*Neural Discrete Representation Learning*" by Aaron van den Oord, Oriol Vinyals and Koray Kavukcuoglu

G2: Variational Autoencoder

Presented on **22.05.2024** by Anastasiia Korzhylova, Ivan Shishkin, Ramneek Agnihotri and Rodi Mehi

Table of contents

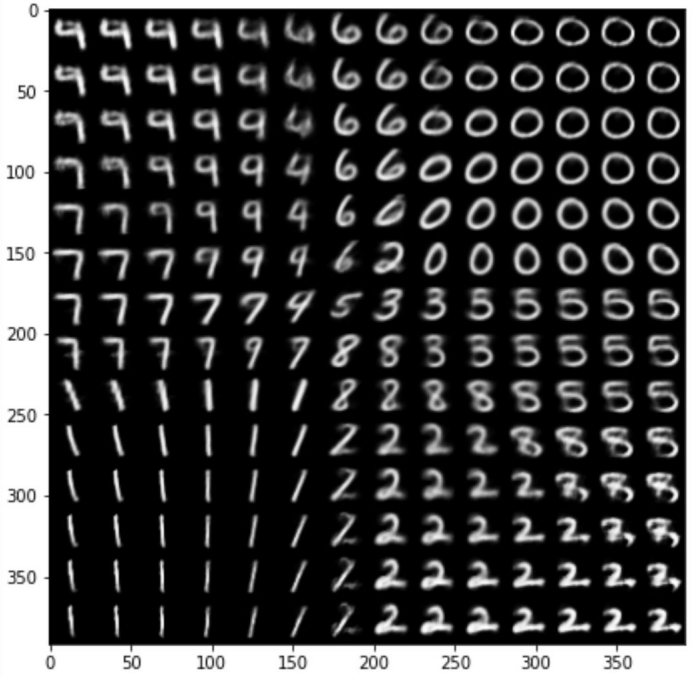
01

Introduction & Motivation

Introduction: What are VAEs useful for?

https://gaussian37.github.io/deep-learning-chollet-8-4/

Generating new data based on previously encoded data

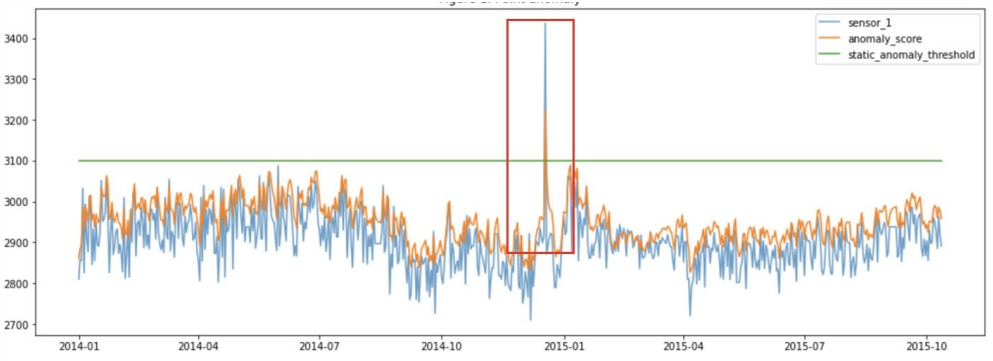


https://arxiv.org/html/2402.12465v1

Introduction: What are VAEs useful for?

- **Anomaly Detection in Network Intruder Detection (Google):**

  if a network packet or sequence of packets exhibits characteristics that differ significantly from the learned normal patterns, it would trigger an alert for potential intrusion.

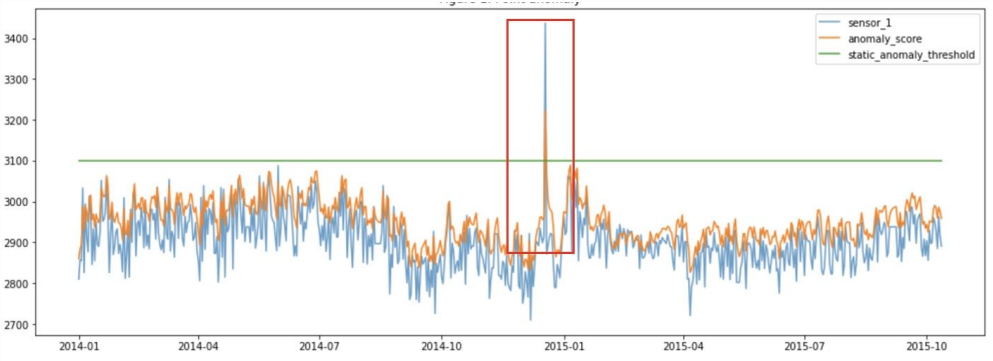


https://developer.ibm.com/developer/default/learningpaths/

Introduction: What are VAEs useful for?

- **Anomaly Detection in Financial Transactions (Uber):**

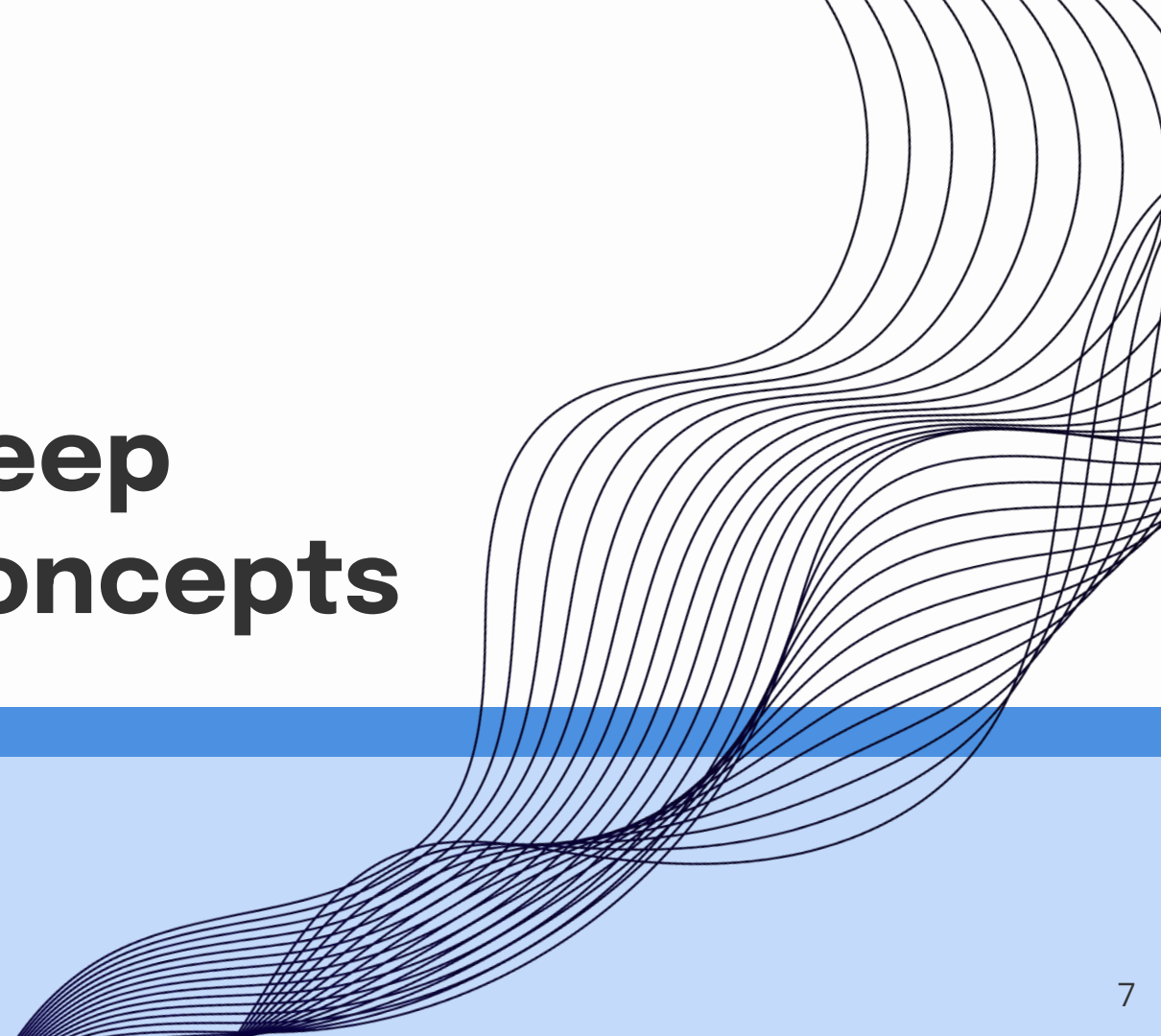
  if the reconstruction error for a given transaction is high or if its latent space representation deviates significantly from the learned distribution of normal transactions, it would be flagged as an anomaly or potential fraud.



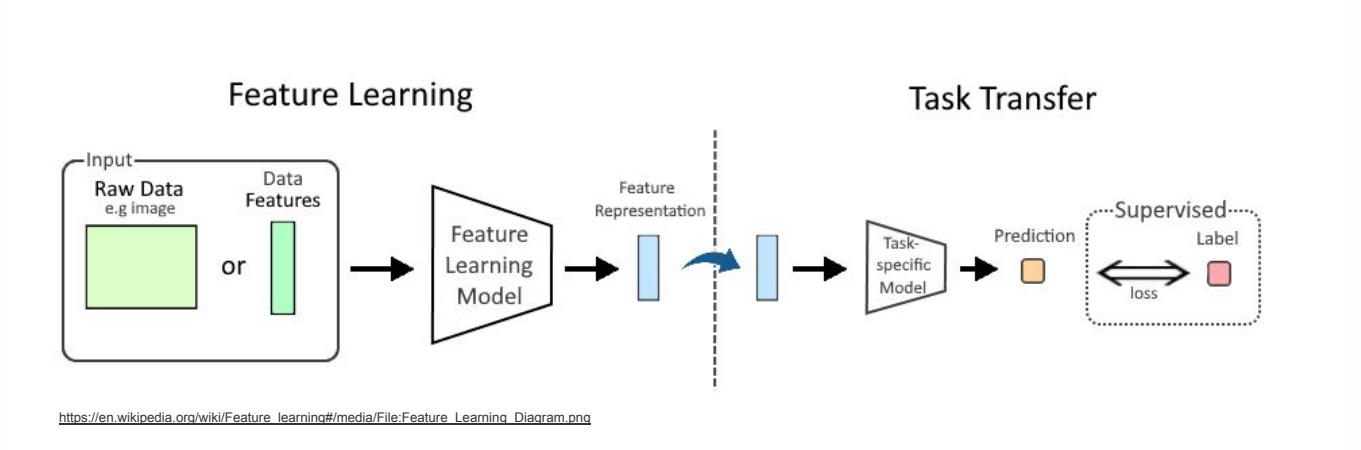
https://developer.ibm.com/developer/default/learningpaths/

02

Relevant Deep Learning concepts



Was ist Representation Learning?



https://en.wikipedia.org/wiki/Feature_learning#/media/File:Feature_Learning_Diagram.png

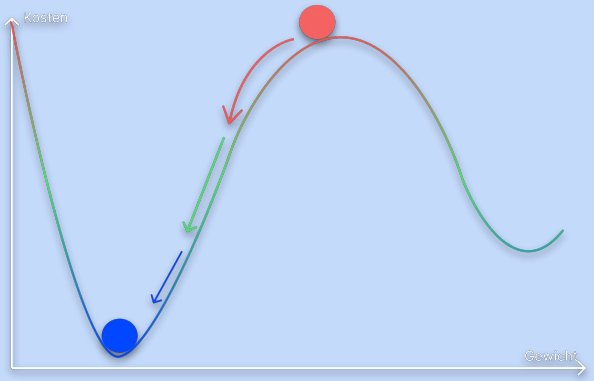
- Learning a representation y = f(x) from an input object (training data) x
- Representation can emphasize different features
- Small representation, easier to understand/process
- Useful for classification, clustering, etc

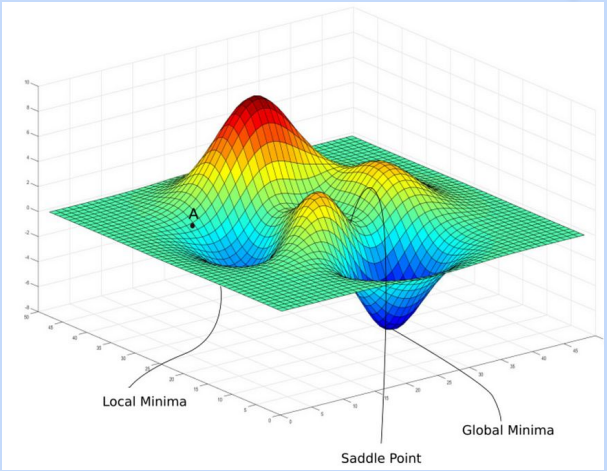
Gradient Descent

Algorithm:
After each forward iteration of the neural network, update each parameter until convergence is achieved

$$\theta_j := \theta_j - \eta \frac{\partial}{\partial \theta_j} L(\theta_0, \theta_1, ..., \theta_n)$$

Use **backpropagation** to determine the gradient of the loss function





https://www.researchgate.net/figure/Fig-7-Parameters-of-the-gradient-descent-variation-optimization-Gradient-Descent_fig6_350567223

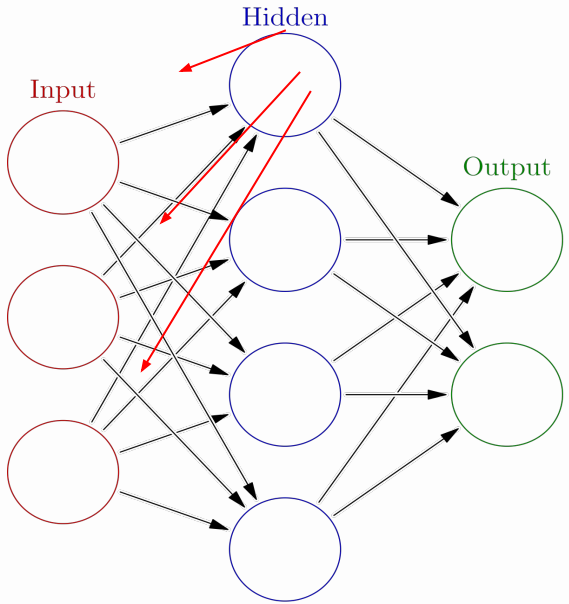
Backpropagation

Objective: (efficient) calculation of the direction of the largest ascent (descent) of the loss function

$$\frac{\partial}{\partial \theta_j} L(\theta_0, \theta_1, ..., \theta_n)$$

After each forward iteration, the network is traversed backwards

The largest ascent is calculated by repeatedly applying the chain rule



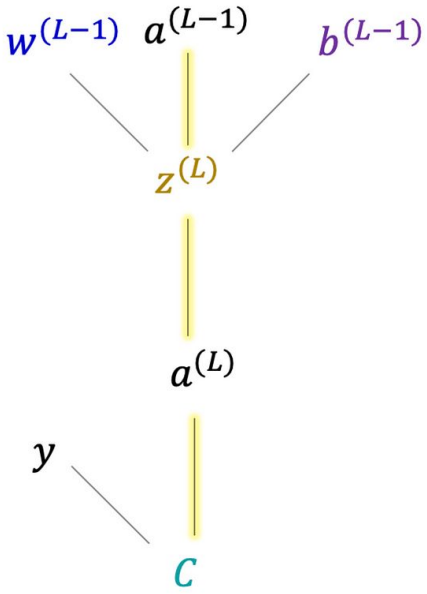
https://www.researchgate.net/figure/Representation-of-interconnected-group-of-nodes-in-an-artificial-neuronal-network-where_fig1_351536133

Backpropagation

What we want: $\frac{\partial C}{\partial a^{(L-1)}}$

Chain rule

$$\frac{\partial C}{\partial a^{(L-1)}} = \frac{\partial z^{(L)}}{\partial a^{(L-1)}} \frac{\partial a^{(L)}}{\partial z^{(L)}} \frac{\partial C}{\partial a^{(L)}}$$



https://towardsdatascience.com/the-maths-behind-back-propagation-cf6714736abf

a(L): Output of the neuron

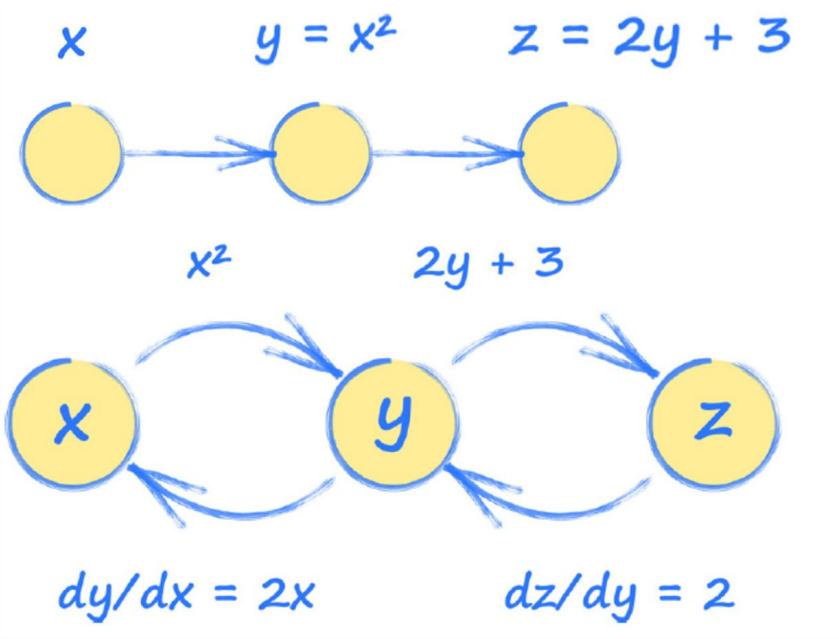
z(L): Output before function

b(L-1), a(L-1), W(L-1) Bias, previous neuron, weight

y label of dataset

C is the Cost function

Backpropagation: example



https://pt.linkedin.com/pulse/afinal-o-que-%C3%A9-essa-tal-de-inteligência-artifical-e-quais-rafael-s

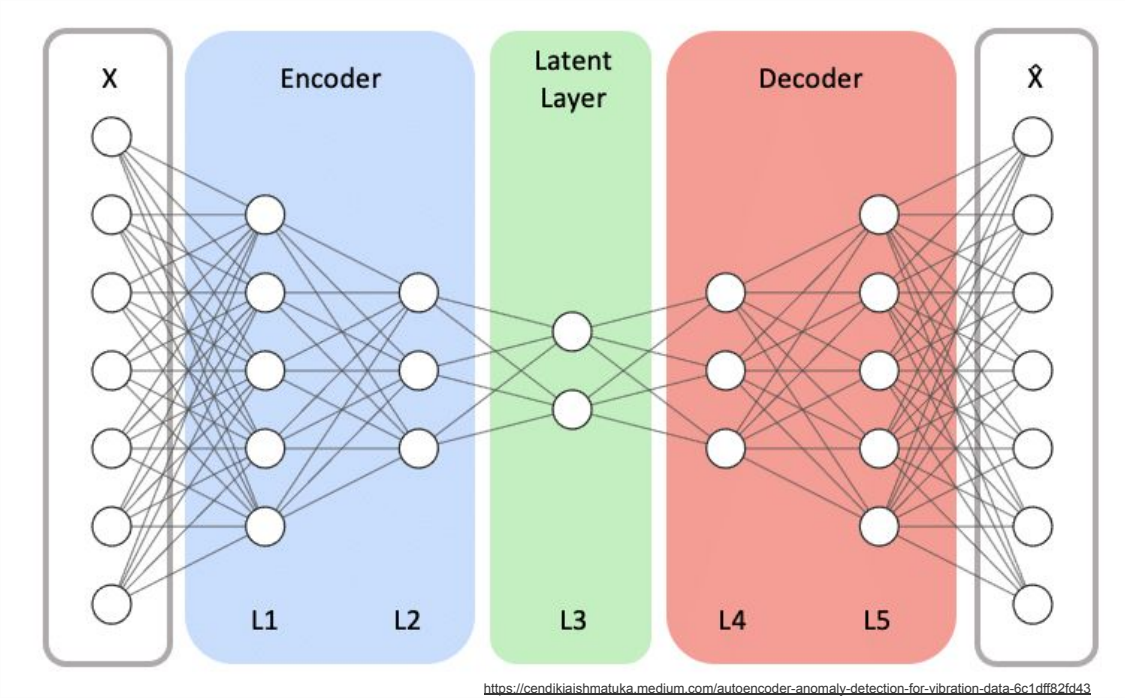
Autoencoder: Architecture and Use Cases

Useful for:

- Feature detection
- Anomaly detection
- Face recognition

1. Compare input and output using a loss function

2. Minimize loss



https://cendikiaishmatuka.medium.com/autoencoder-anomaly-detection-for-vibration-data-6c1dff82fd43

03

VAE: Variational Autoencoder

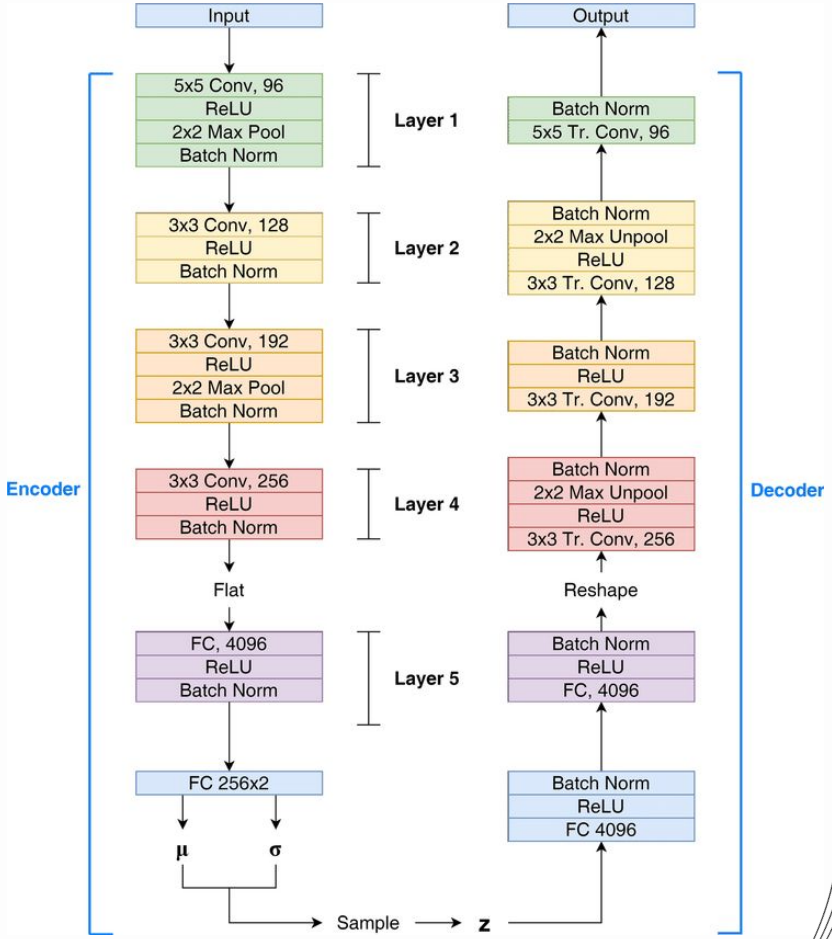
VAE Architecture

Encoder:

- downsamples and transforms input data into a lower-dimensional representation in latent space
- consists of convolutional, pooling, flattering and dense layers

Decoder:

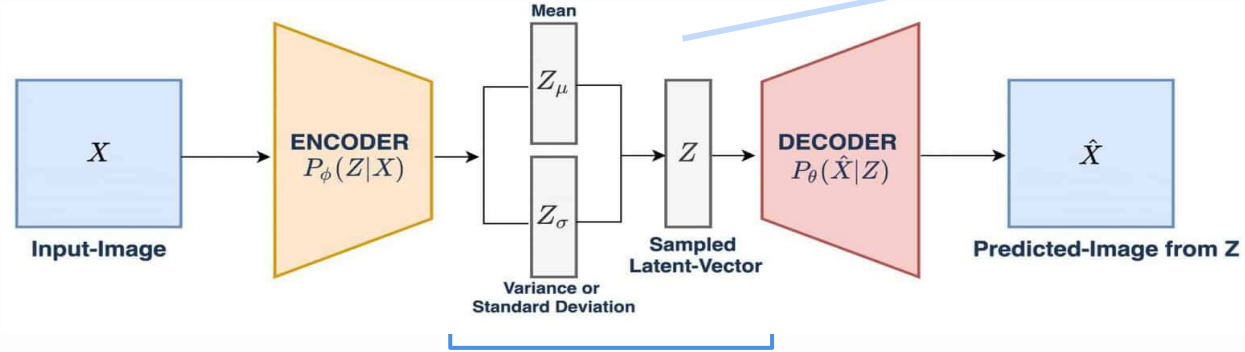
- reconstructs original input data from a point from the latent space by gradual upsampling
- consists of dense, reshaping and transposed convolutional layers (deconvolution)

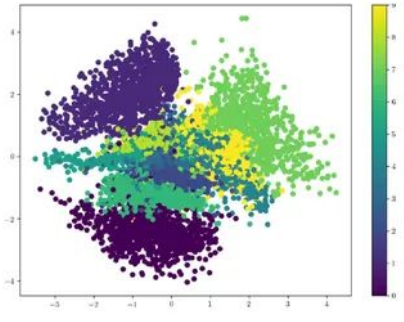


https://www.researchgate.net/publication/357911187/figure/fig5/AS:1134446157803524@1647484449716/The-encoder-decoder-architecture-for-the-Variational-Auto-Encoder-VAE-experiments.png

VAE Architecture

https://learnopencv.com/wp-content/uploads/2020/11/vae-diagram-1-2048x1126.jpg





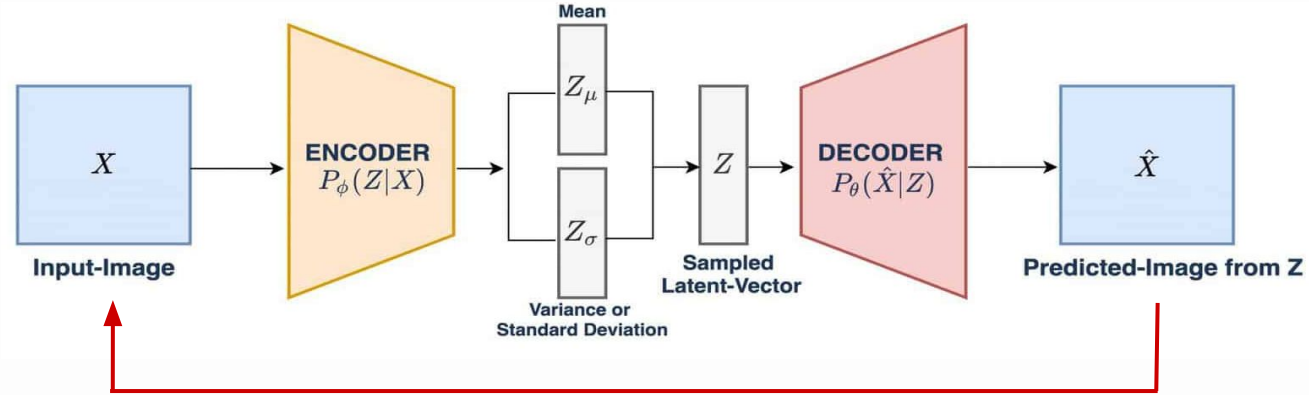
https://av-eks-lekhak.s3.amazonaws.com/media/__sized__/article_images/g4_thH4l3P-thumbnail_webp-600x300.webp

Latent space

- a lower-dimensional representation of the input data
- serves as a bottleneck, capturing the essential features of the input data in a compact form
- mean and variance parameters output by the encoder are used to sample points in the latent space using **reparameterization** that ensures differentiability during training

VAE Architecture

https://learnopencv.com/wp-content/uploads/2020/11/vae-diagram-1-2048x1126.jpg



Reconstruction Loss + KL Divergence

Measures the difference between the input data and its reconstruction generated by the decoder.

Measures the divergence between the distribution of latent variables produced by the encoder and a chosen prior distribution.

VAE: the prior p(z) and the latent variable z

$$p_{\theta}(\mathbf{z}|\mathbf{x}) = p_{\theta}(\mathbf{x}|\mathbf{z})p_{\theta}(\mathbf{z})/p_{\theta}(\mathbf{x})$$

The (intractable) posterior probability (encoder) is dependant on the prior

Latent variable z:

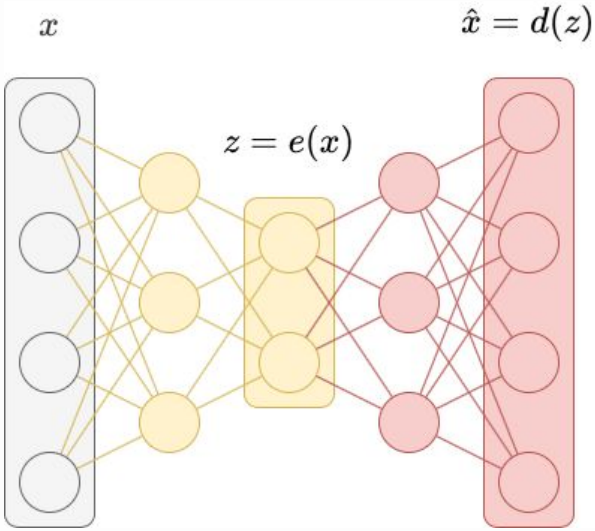
- the latent variable is a vector from the hidden space in the variational autoencoder
- Is it a smaller vector than the input vectors
- The encoder, encodes inputs into the latent space as a latent vector

The prior p(z):

- The prior is the assumed distribution about the latent variable z
- Data is sampled from prior, and then decoded (using a neural network)
- A distribution selected a priori,
- Often: the normal distribution with mean 0, and variance 1 selected

MNIST-Dataset

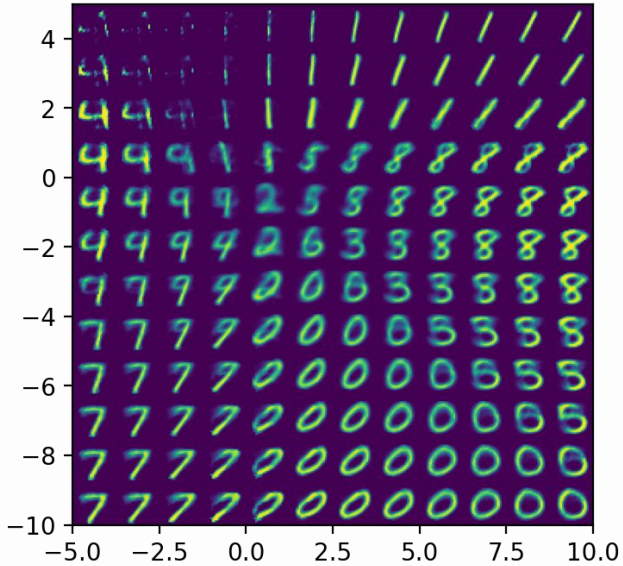
- X = {x1, x2, …, xn}, where x = 28 x 28 matrix, flattened to 784-dim vector.
- In real machine learning problem, datapoints contain **thousands of dimensions**.
- Instead of creating a fully connected neural network, we **reduce the dimensions** and create a **“latent” vector** z.
- From this “latent” vector we should be able to recreate the images



https://avandekleut.github.io/vae/

MNIST-Dataset

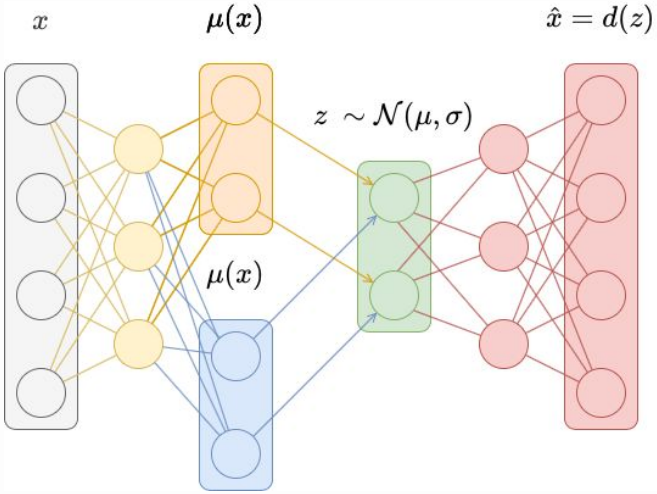
- Purpose of generative modeling is to **create new data points**
- Using autoencoding to calculate the latent vector z **leaves gaps in the latent space**
- This makes it hard to create unseen images
- The latent space z is disjoint and non-continuous



https://avandekleut.github.io/vae/

MNIST-Dataset

- Instead of mapping our images to a latent vector z, we map to a distribution to a **mean vector** $\mu(x)$ and a vector of **standard deviations** $\sigma(x)$
- These parametrize a diagonal **Gaussian distribution** $\varepsilon \sim \mathcal{N}(\mu, \sigma)$
- From this distribution we sample out latent vector z



https://avandekleut.github.io/vae/

Lower Bound of p(x)

- The probability p(x) is the intractable distribution of images
- Instead of maximizing log(p(x)) **maximize its lower bound:**

$$\mathrm{ELBO}(\theta, \phi) = \mathbb{E}_{q_\phi(z|x)}[\log p_\theta(x|z)] - \mathrm{KL}(q_\phi(z|x)||p(z))$$

- Our expected value E is calculated by using the log-likelihood of all data points
- With the **Kullback Leibler Divergence** we measure the difference between our prior p(z) and our distribution q(z|x) (which is an approximation of p(z|x))

Digression: Log-Likelihood

- **Numerical Stability:**

  Probabilities can take very small values, which can lead to numerical problems. The logarithm transforms these small values into larger, more manageable numbers.

- **Simpler Calculation:**

  The log-likelihood allows us to transform products of probabilities into sums of log-probabilities:

$$\log(p(x_1 \mid \theta) \cdot p(x_2 \mid \theta) \cdot \ldots \cdot p(x_n \mid \theta)) = \log p(x_1 \mid \theta) + \log p(x_2 \mid \theta) + \ldots + \log p(x_n \mid \theta)$$

  This is particularly useful when dealing with large datasets.

Digression: Kullback Leibler Divergence

- The **Kullback–Leibler (KL) Divergence** is a measure of how one probability distribution diverges from the second, expected probability distribution.

$$KL(P \parallel Q) = \sum_{x} P(x) \log \left(\frac{P(x)}{Q(x)} \right)$$

- Here, **P** is the true distribution (or target distribution) and **Q** is the approximated distribution (or model distribution).

Kullback Leibler Loss

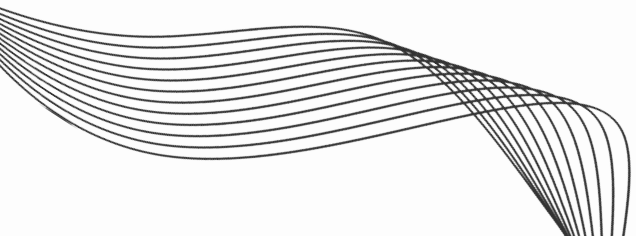
- encourages the **approximation of the posterior distribution q(z|x) of the latent variables to match a specified prior distribution**
- Minimizing the KL divergence loss prevents the posterior distribution from deviating significantly from the prior distribution

$$\mathcal{L}_{\mathrm{KL}} = \mathrm{KL}(q_\phi(z|x)||p(z))$$

Reconstruction Loss

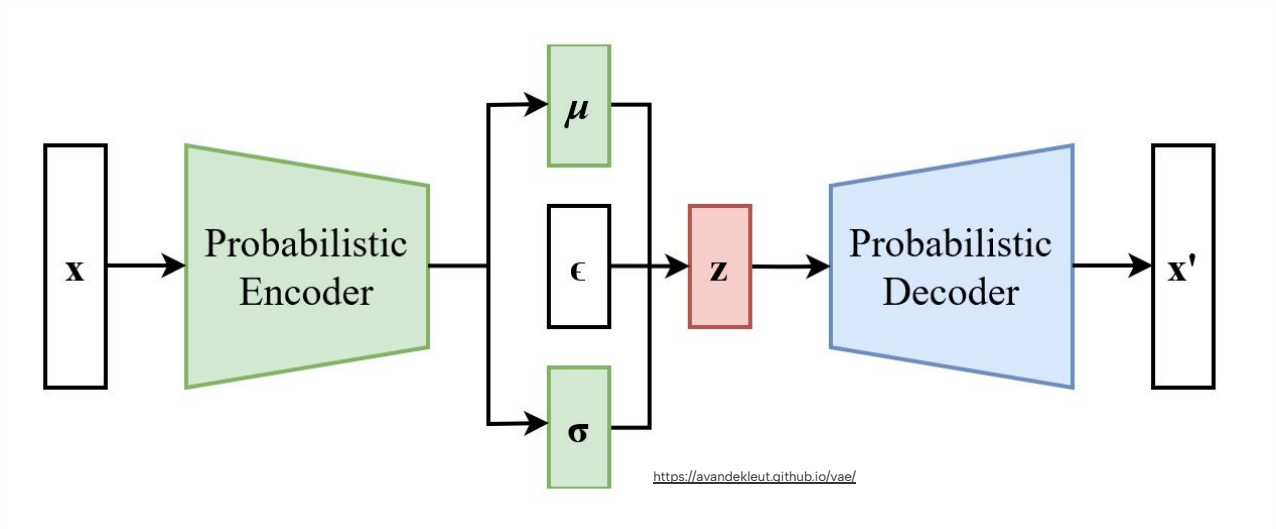
- penalizes the **discrepancy between the original input data and its reconstruction**
- Minimizing the reconstruction loss acts as a regularization term, preventing overfitting

$$\mathcal{L}_{\text{rec}} = -\mathbb{E}_{q_\phi(z|x)}[\log p_\theta(x|z)]$$



Reparameterization Trick

- Because of our parameter ϕ and θ we **can't use the method of the gradient descent** with the term of our expected value
- We use a standard random value $\varepsilon \sim \mathcal{N}(0, \mathrm{I})$: $z = \mu + \sigma \odot \varepsilon$



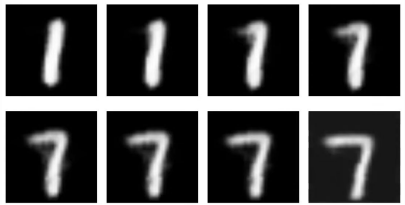
04

VQ-VAE vs. VAE

Discrete vs. Continuous Latent Codes

Standard VAE

Standard VAEs typically use continuous latent variables, often modeled as Gaussian distributions. These continuous latent variables capture the **variability** in the input data.



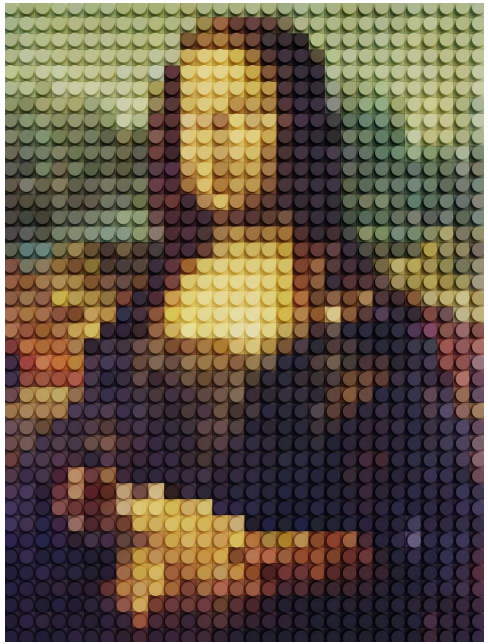
https://www.youtube.com/watch?v=1ZHzAOutcnw

VQ-VAE

VQ-VAE employs discrete latent variables. Instead of representing the latent space as a continuous distribution, VQ-VAE discretizes the latent space into a finite set of discrete codes, saved in the **codebook**. Each code corresponds to a specific **category** or prototype in the latent space.

https://deepai.org/machine-learning-model/text2img
(prompt "person with a cat head")

Learned Prior vs. Static Prior:



https://miro.medium.com/v2/resize:fit:1100/format:webp/1*e3sHa808xu5f0kdqbpYyRQ.jpeg

Standard VAE

Standard VAEs often assume a simple prior distribution, such as a standard **Gaussian distribution**, for the latent variables. This prior is usually **static** and does not change during training.

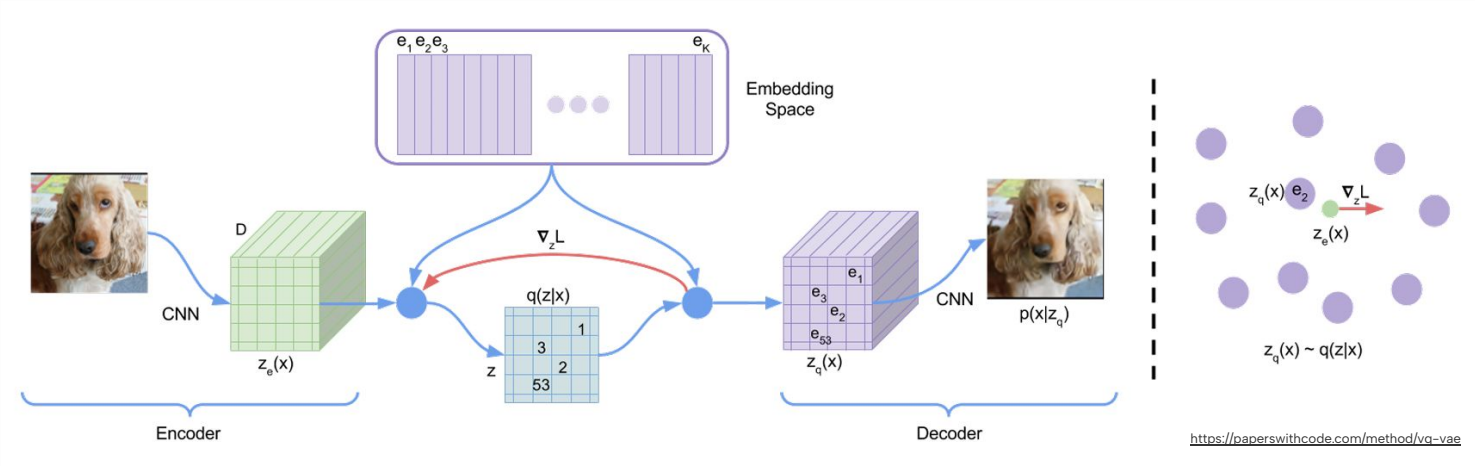
VQ-VAE

VQ-VAE learns the prior distribution over the discrete latent space **dynamically**. After training the VQ-VAE, an autoregressive model is trained to learn the **conditional probability distribution** of each discrete code given its context.

Vector quantization (VQ)

VQ is a process used to **map continuous input vectors** from the encoder **to discrete latent codes** for the decoder to minimize the codebook loss.

1. The codebook consists of a set of learnable embedding vectors representing discrete codes.
2. The input vector is compared to each embedding vector in the codebook, and the nearest embedding vector (in terms of **Euclidean distance**, for instance) is selected as the quantized representation.



https://paperswithcode.com/method/vq-vae

Vector quantization (VQ)

Challenges

Discretization Error

Loss of input information during VQ. Reduced by commitment loss and codebook loss.

Codebook Collapse

Not all embedding vectors in the codebook might get utilized effectively. Solved by EMA updates and Random Restart.

Training Instability

Caused by updating both the encoder and the codebook. Prevented by stop-gradient or training in separate steps.

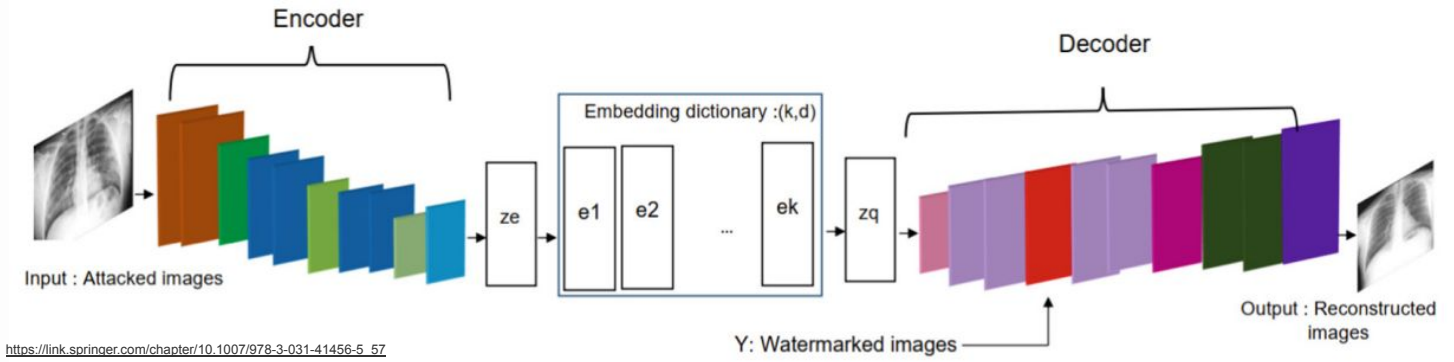
Overhead

Managing a large codebook is computationally expensive. Use hierarchical VQ or efficient searching (KD-trees, ANN) instead.

Loss function

$$L = \log p(x|zq(x)) + \|sg[ze(x)] - e\|^2_2 + \beta\|ze(x) - sg[e]\|^2_2$$

Reconstruction loss **VQ loss** **Commitment loss**



https://link.springer.com/chapter/10.1007/978-3-031-41456-5_57

Reconstruction loss $\log p(x|zq(x))$

⇒ Measures the **difference** between the original data and the decoder output

⇒ The **goal** is to make the output of the decoder **as close as possible** to the original input

⇒ This is often done by **maximizing** the log-likelihood of the data, which is equivalent to **minimizing** the negative log-likelihood represented by this formula

$$\log P(x \mid z) = -\frac{1}{2}\left[\log(|\Sigma|) + \mathrm{klog}(2\pi) + (\mathbf{x}-\mu)^T(\mathbf{x}-\mu)\right]$$

first two terms are constant with respect to μ

$$-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T(\mathbf{x}-\boldsymbol{\mu})$$

MSE

⇒ **x** - the input data

⇒ $\mathbf{z_q(x)}$ - discrete latent variables produced by the encoder

VQ (Codebook) loss & Commitment loss

⇒ **VQ** : encourages the discrete latent variables to move towards the encoder outputs

⇒ **Comm.** : encourages the encoder output to commit as much as possible to its closest codebook vector

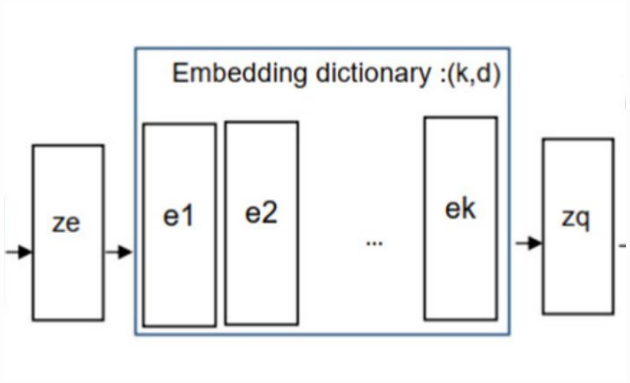
$$\underbrace{\|sg[ze(x)] - e\|^2_2}_{\text{VQ loss}} + \underbrace{\beta\|ze(x) - sg[e]\|^2_2}_{\text{Commitment loss}} \quad \text{vs.} \quad ||z_e(x) - e||^2$$

⇒ ***β*** – hyperparameter, controls the **weight** of the comm. loss in the overall loss function

⇒ **sg[]** – stop gradient operator

⇒ **$z_e(x)$** – the encoder output

⇒ **e** – the embeddings



https://link.springer.com/chapter/10.1007/978-3-031-41456-5_57

Powerful decoder

⇒ Decoder that can reconstruct the input data from the latent space with **high fidelity**

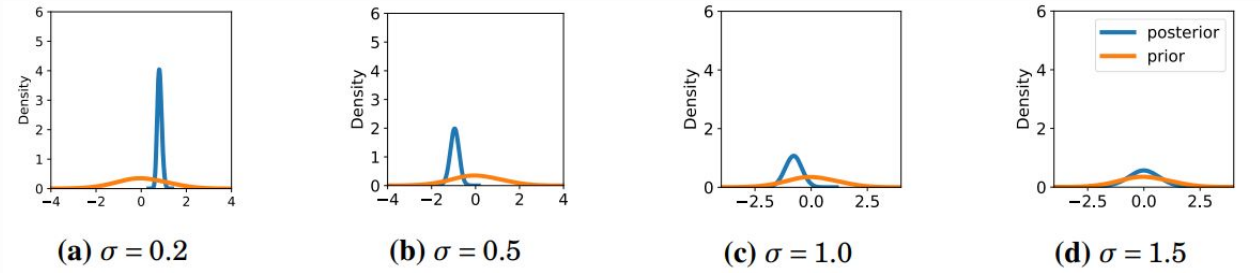
⇒ Decoder that is capable of modeling the data density **without relying heavily** on the latent variables

⇒ **Problem**: if the decoder is too powerful, it can lead to the following issue…

Posterior collapse

⇒ **Phenomenon** where the model ignores the latent variables and relies solely on the decoder to model the data

⇒ **Result** is a trivial posterior that collapses to the prior

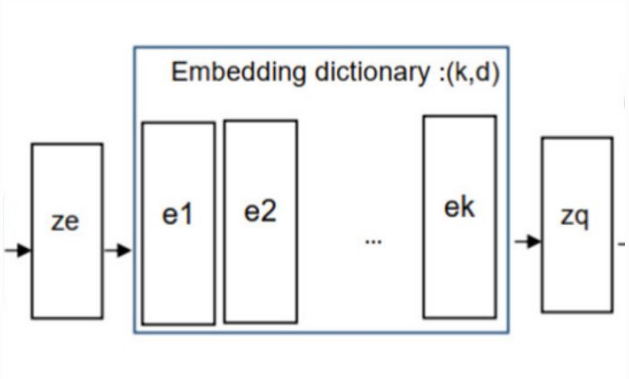


https://proceedings.neurips.cc/paper_files/paper/2021/file/2b6921f2c64dee16ba21ebf17f3c2c92-Paper.pdf

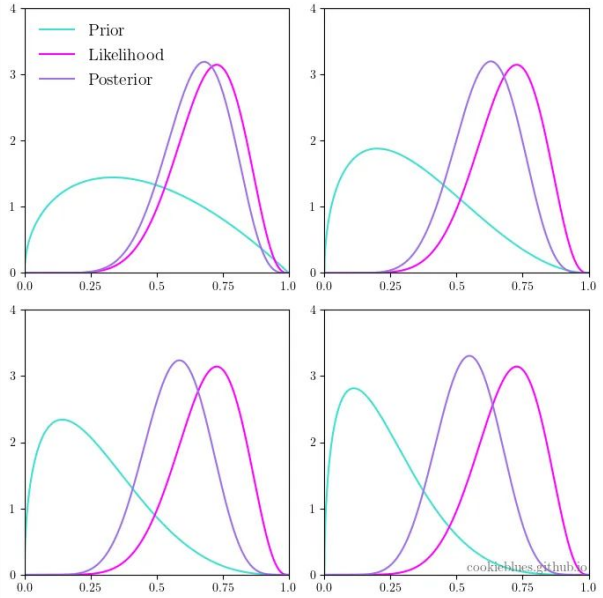
Posterior collapse avoidance

⇒ VQ-VAE avoids the issue of “posterior collapse” by using discrete latent variables instead of continuous ones

⇒ The discrete representation (the codebook vectors) forces the model to use the latent variables more effectively, as it cannot rely on a continuous space to encode the data



https://link.springer.com/chapter/10.1007/978-3-031-41456-5_57



https://towardsdatascience.com/what-is-bayesian-inference-4eda9f9e20a6

Experiment 1: Images

https://arxiv.org/pdf/1711.00937

Figure 2: Left: ImageNet 128x128x3 images, right: reconstructions from a VQ-VAE with a 32x32x1 latent space, with K=512.

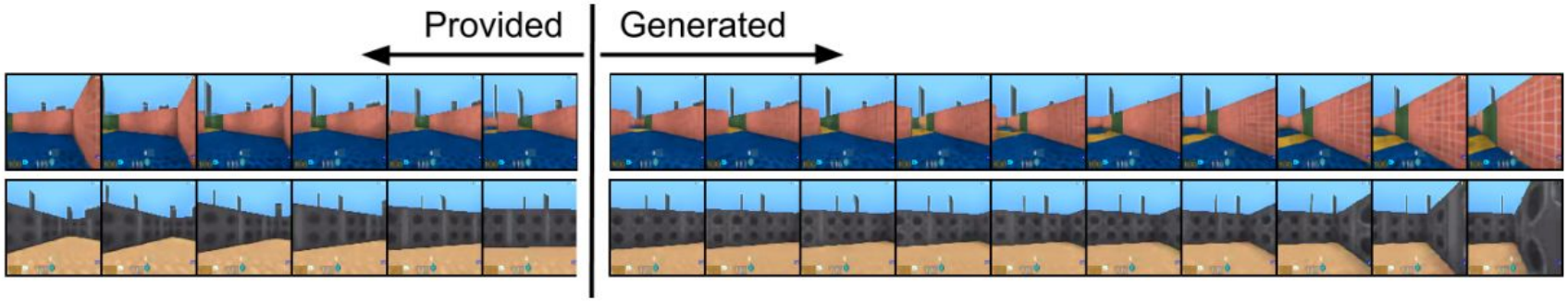
⇒ The reconstructions look only **slightly blurrier** than the originals

⇒ Images were modeled by learning a powerful prior (**PixelCNN**) over z. This allows to not only greatly speed up training and sampling, but also to use the PixelCNNs capacity to capture the global structure instead of the low-level statistics of images.

Experiment 2: Video

⇒ Generation of the video sequence is done **purely** in the latent space, **without** the need to generate the actual images themselves

⇒ Each image is then created by **mapping** the latents with a deterministic decoder to the pixel space. Latents are generated using **only** the prior model



https://arxiv.org/pdf/1711.00937

Figure 7: First 6 frames are provided to the model, following frames are generated conditioned on an action. Top: repeated action "move forward", bottom: repeated action "move right".

Thanks!

Do you have any questions?

Mastering mathematical communication

Clarity and precision

The importance of articulating mathematical ideas with precision

Justification and proof

Justify mathematical solutions, providing proof to validate conclusions

Active listening

Engage with peers' mathematical explanations and provide feedback

Visual representation

The use of diagrams, graphs, and charts to supplement written explanations

Logical organization

Structure mathematical reasoning coherently in a logical sequence

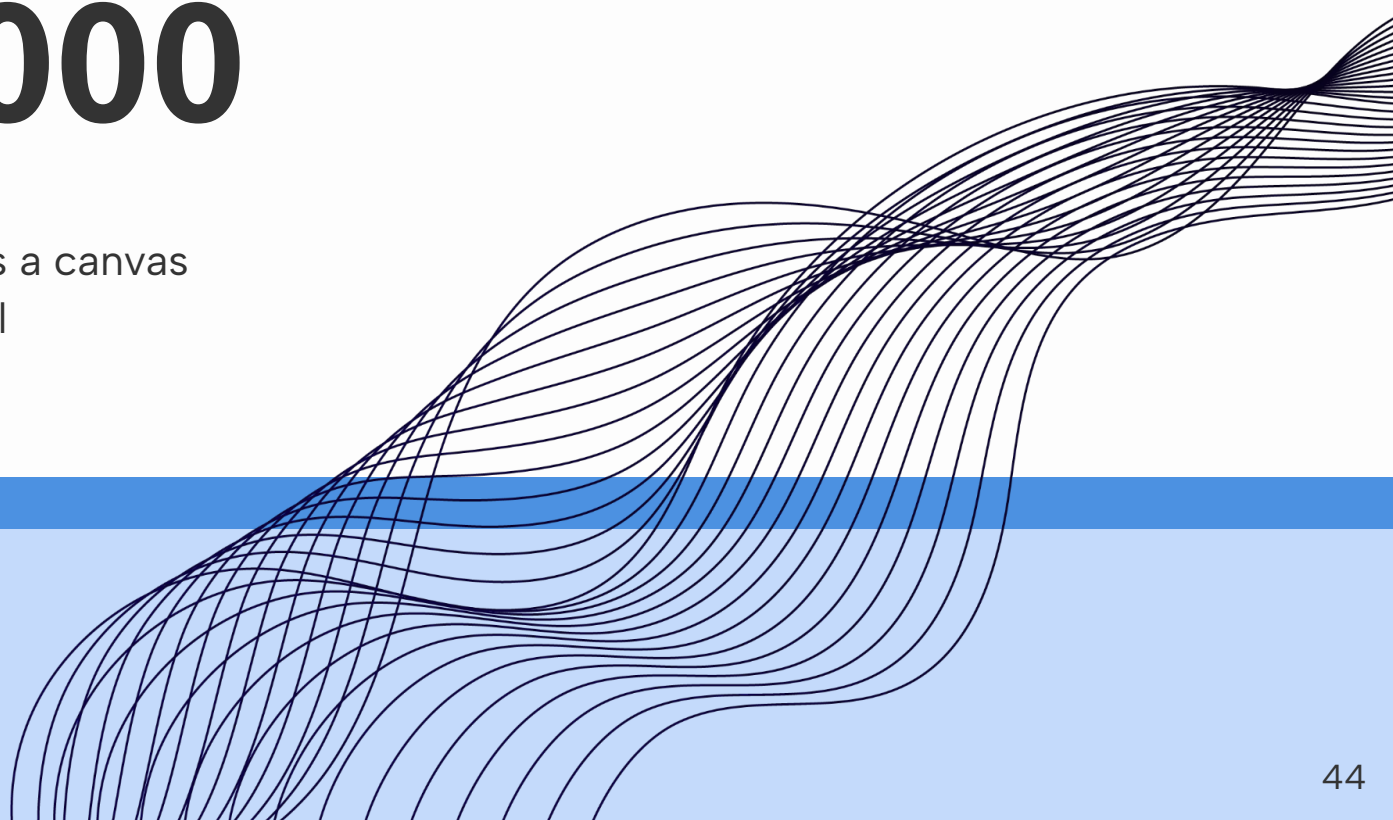
Adaptability

Adapt your communication style to different audiences to convey ideas

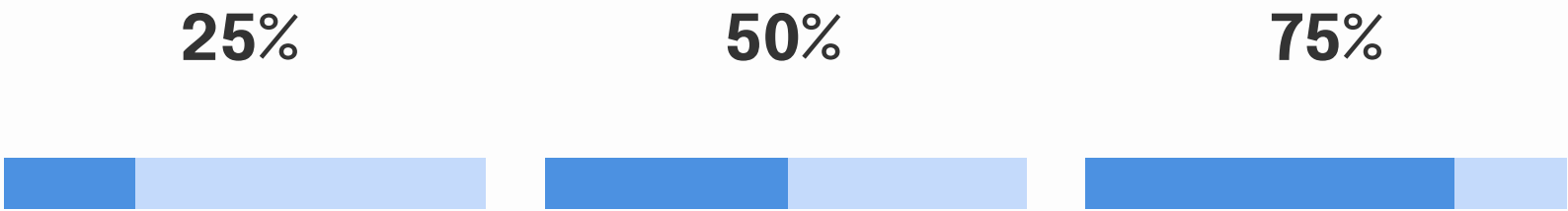

Mathematics is the tool that turns curiosity into discovery

123,000

Numbers unfold as a canvas of infinite potential



Let’s use some percentages!



Theory

25% of the math curriculum is dedicated to theoretical knowledge

Foundation

50% of the curriculum is spent creating a solid understanding of maths

Exploration

75% of the curriculum cultivates advanced analytical skills

Milestones in mathematics

Euclid publishes "Elements"

300 BCE

Gödel's Incompleteness Theorem

1931

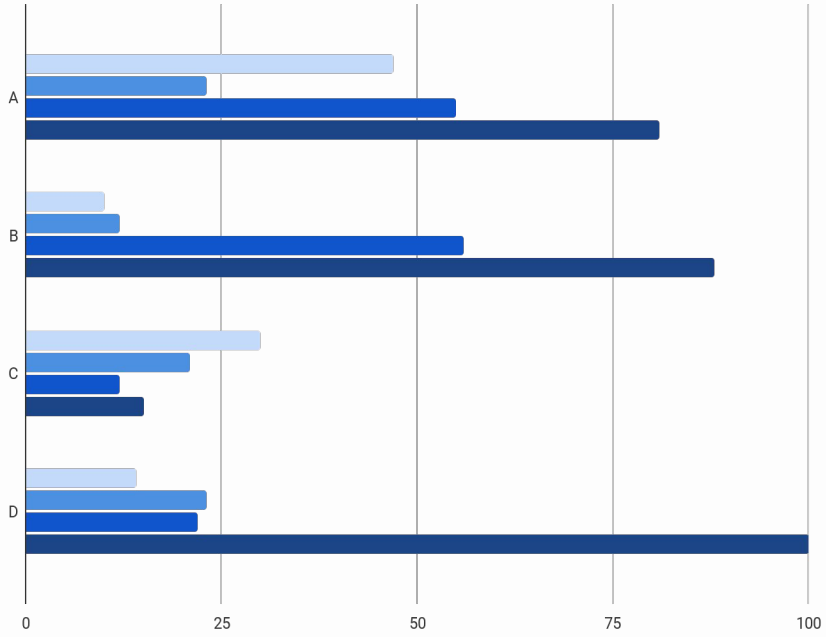
17th century

The development of calculus

1994

Proof of Fermat's Last Theorem

Display concepts visually



Concept 1

Briefly describe the concept here

Concept 2

Briefly describe the concept here

Concept 3

Briefly describe the concept here

Concept 4

Briefly describe the concept here

Follow the link in the graph to modify its data and then paste the new one here. **For more info, click here**

Rubric

Criteria	Exceptional (4)	Proficient (3)	Basic (2)	Limited (1)
Conceptual understanding	Deep, insightful mastery of concepts	Solid understanding and application	Basic grasp, occasional struggles	Limited understanding, frequent errors
Problem-solving skills	Exceptional, creative problem-solving	Strong skills with various strategies	Basic skills, occasional struggles	Limited ability, frequent challenges
Communication of ideas	Clear, precise communication	Effective expression of mathematical ideas	Adequate but may lack precision	Struggles to communicate clearly
Accuracy of calculations	Exceptional accuracy in computations	High accuracy in routine calculations	Basic accuracy, occasional errors	Frequent errors in calculations

Short answer questions

The students should read the following questions and answer them in their paper. These questions require students to demonstrate their understanding of key concepts

1	**How would you write an algebraic expression for "three times the sum of a number and 5"?**
2	**Can you explain the distinction between the area and perimeter of a shape? Provide an example for each**
3	**What is the result when you add 1/3 and 1/4? Express your answer as a simplified fraction**

Real-life application

Divide the class into small groups of 3-5 students. Assign each group a real-life application for which they will have to use math concepts to solve

Topics

Budgeting for a trip

Consider transportation, accommodation, meals and other activities

Personal finance

Include expenses like rent, utilities, groceries and entertainment

Architectural design

Design a floor plan, calculating the area of each room with proper proportions

Epidemiology

Using statistical methods, analyze data on infection and recovery rates

Mathematical investigation

Students will have to delve into a specific mathematical concept or problem, explore it in-depth, and present their findings

1. **Choose a number sequence:** Ask each student to select a number sequence that interests them
2. **Investigate the pattern:** Instruct them to explore the pattern within their chosen number sequence
3. **Apply mathematical concepts:** They should apply relevant mathematical concepts to their investigation
4. **Solve problems:** They have to create and solve problems related to their chosen sequence
5. **Create a portfolio:** Instead of a traditional essay, ask them to compile their findings into a portfolio and present it

Conclusions

Summarize key concepts

Recap the main mathematical concepts covered during the lesson

Real-world applications

Emphasize the practical relevance of the concepts and skills of the lesson

Preview next steps

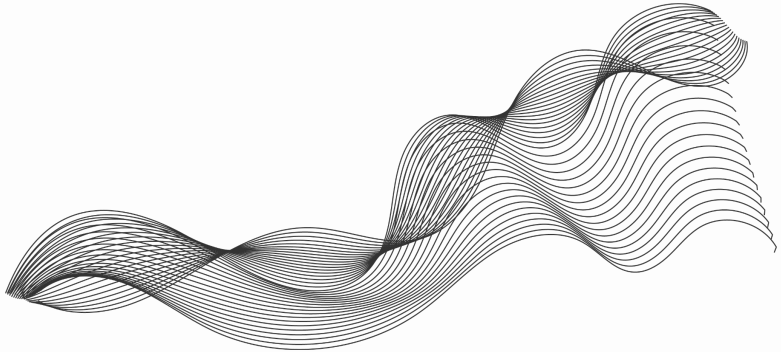
Provide a glimpse into the upcoming lessons for continuity of learning

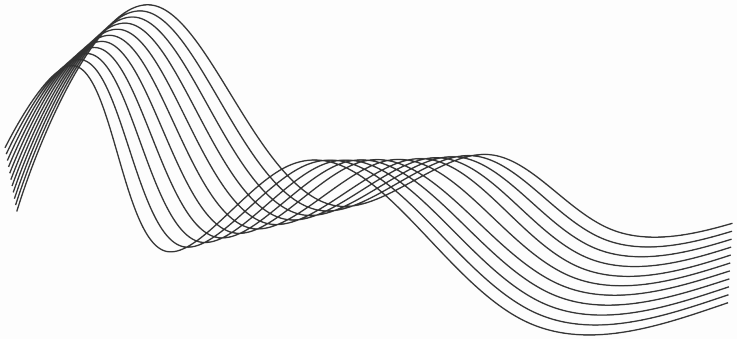
Alternative resources

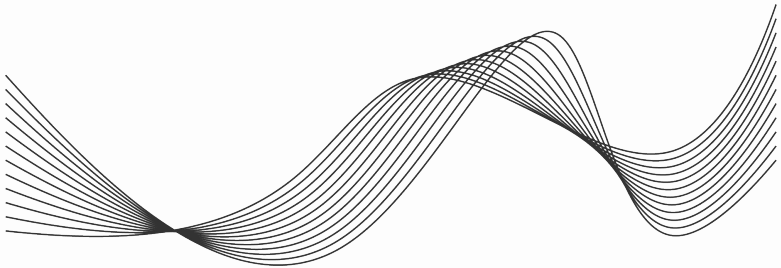
Here’s an assortment of alternative resources whose style fits that of this template:

Vectors

- Lineal shapes landing page I
- Lineal shapes landing page II







Resources

Did you like the resources used in this template? Get them on these websites:

Vectors

- Lineal shapes landing page template

Photos

- Medium shot people studying math I
- Medium shot people studying math II

Instructions for use

If you have a free account, in order to use this template, you must credit **Slidesgo** by keeping the **Thanks** slide. Please refer to the next slide to read the instructions for premium users.

As a Free user, you are allowed to:

- Modify this template.
- Use it for both personal and commercial projects.

You are not allowed to:

- Sublicense, sell or rent any of Slidesgo Content (or a modified version of Slidesgo Content).
- Distribute Slidesgo Content unless it has been expressly authorized by Slidesgo.
- Include Slidesgo Content in an online or offline database or file.
- Offer Slidesgo templates (or modified versions of Slidesgo templates) for download.
- Acquire the copyright of Slidesgo Content.

For more information about editing slides, please read our FAQs or visit our blog:
https://slidesgo.com/faqs and **https://slidesgo.com/slidesgo-school**

Instructions for use (premium users)

As a Premium user, you can use this template without attributing **Slidesgo** or keeping the "**Thanks**" slide.

You are allowed to:

- Modify this template.
- Use it for both personal and commercial purposes.
- Hide or delete the “Thanks” slide and the mention to Slidesgo in the credits.
- Share this template in an editable format with people who are not part of your team.

You are not allowed to:

- Sublicense, sell or rent this Slidesgo Template (or a modified version of this Slidesgo Template).
- Distribute this Slidesgo Template (or a modified version of this Slidesgo Template) or include it in a database or in any other product or service that offers downloadable images, icons or presentations that may be subject to distribution or resale.
- Use any of the elements that are part of this Slidesgo Template in an isolated and separated way from this Template.
- Register any of the elements that are part of this template as a trademark or logo, or register it as a work in an intellectual property registry or similar.

For more information about editing slides, please read our FAQs or visit our blog:
https://slidesgo.com/faqs and **https://slidesgo.com/slidesgo-school**

Fonts & colors used

This presentation has been made using the following fonts:

Epilogue
(https://fonts.google.com/specimen/Epilogue)

Albert Sans
(https://fonts.google.com/specimen/Albert+Sans)

Storyset

Create your Story with our illustrated concepts. Choose the style you like the most, edit its colors, pick the background and layers you want to show and bring them to life with the animator panel! It will boost your presentation. Check out **how it works**.

Pana

Amico

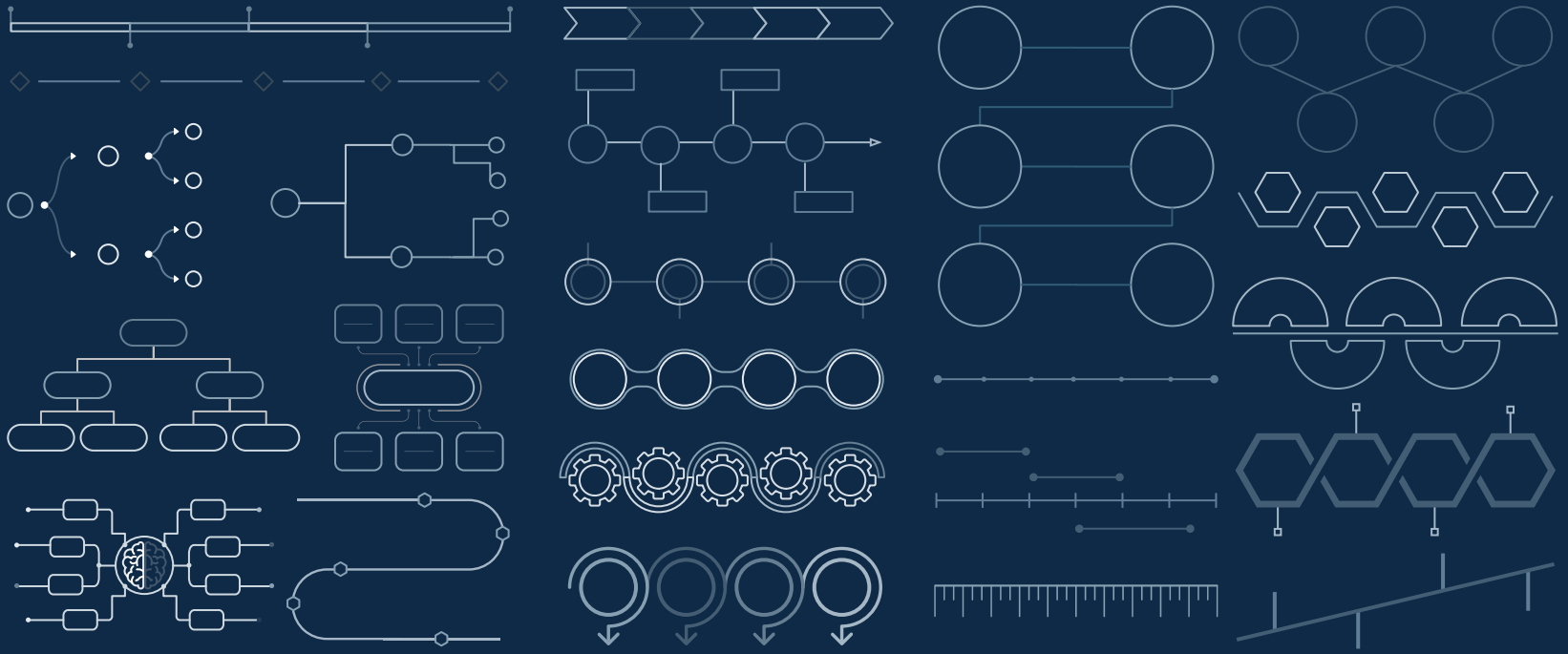
Bro

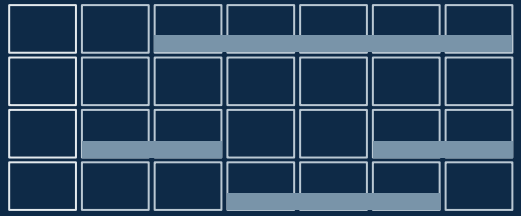
Rafiki

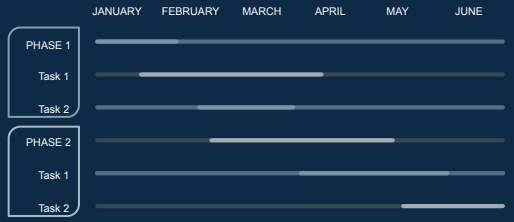
Cuate

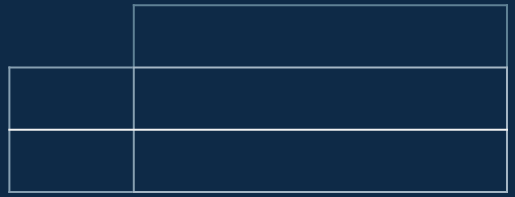
Use our editable graphic resources...

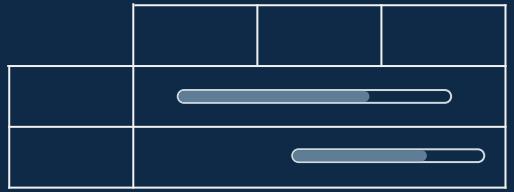
You can easily **resize** these resources without losing quality. To **change the color**, just ungroup the resource and click on the object you want to change. Then, click on the paint bucket and select the color you want. Group the resource again when you're done. You can also look for more **infographics** on Slidesgo.

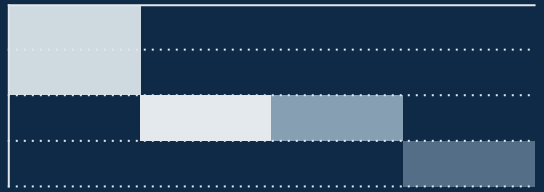


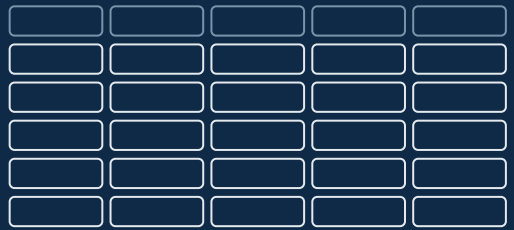


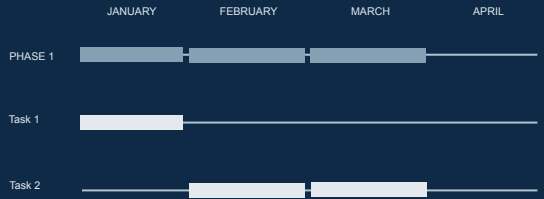

JANUARY
FEBRUARY
MARCH
APRIL
MAY
JUNE
PHASE 1
Task 1
Task 2
PHASE 2
Task 1
Task 2

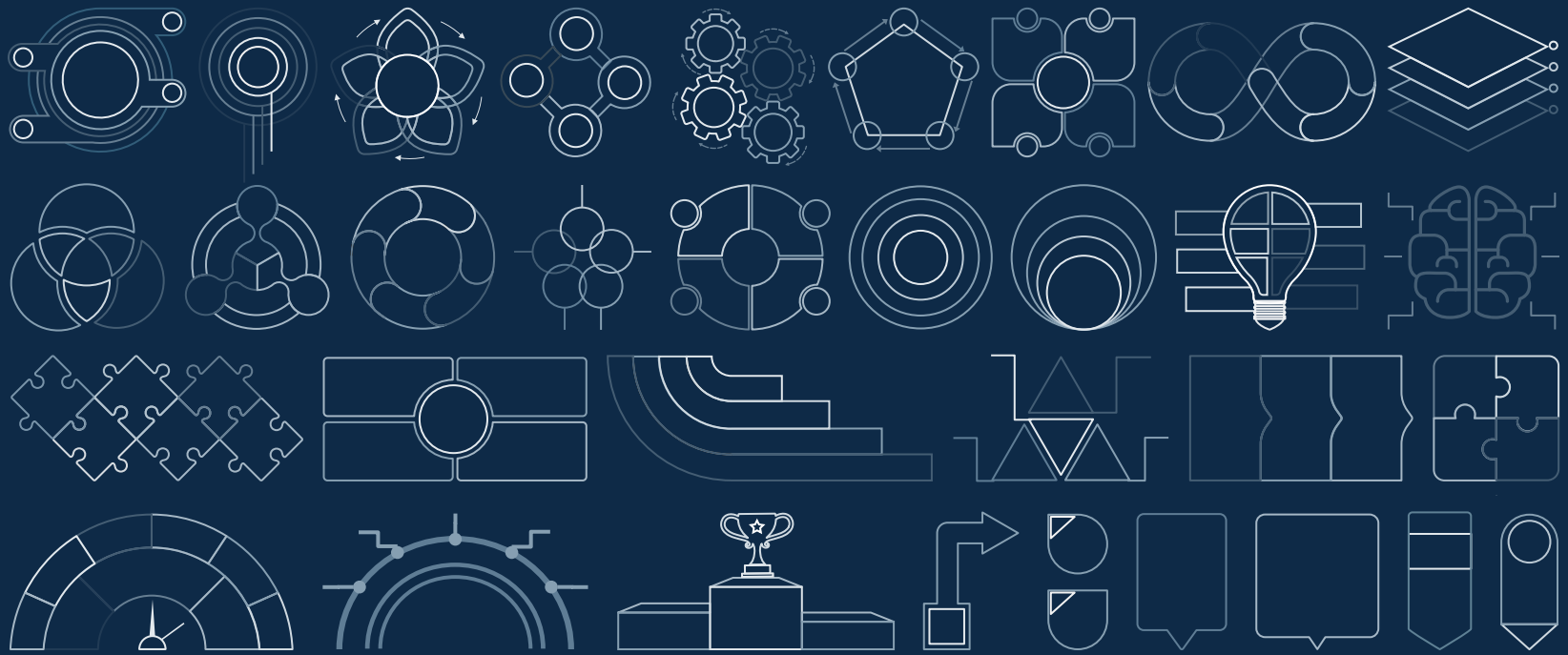


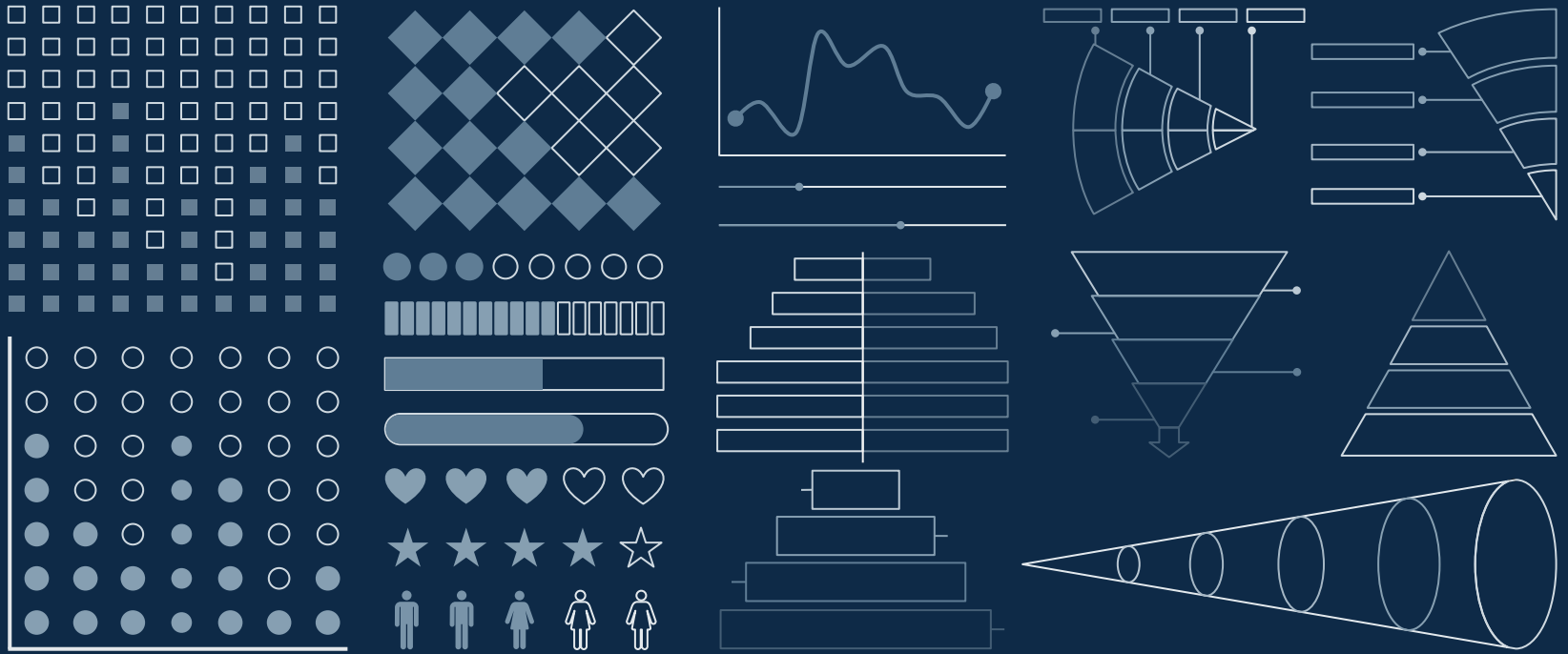







JANUARY
FEBRUARY
MARCH
APRIL
PHASE 1
Task 1
Task 2





...and our sets of editable icons

You can **resize** these icons without losing quality.
You can **change the stroke and fill color**; just select the icon and click on the **paint bucket/pen**.
In Google Slides, you can also use **Flaticon's extension**, allowing you to customize and add even more icons.

Educational Icons

Medical Icons

Business Icons

Teamwork Icons

Help & Support Icons

Avatar Icons

Creative Process Icons

Performing Arts Icons

Nature Icons

SEO & Marketing Icons


slidesgo