UNSUPERVISED LEARNING

IN5400 — Machine Learning for Image Analysis

Ole-Johan Skrede

03.04.2019

University of Oslo

Messages

- Mandatory 2 is ready soon (some technical difficulties)
- Exercise for this week is ready before tomorrow

Outline

- Introduction and motivation
- Repetition / background
  - K nearest neighbours, k-means clustering
  - Principal component analysis
  - Independent component analysis
- t-SNE
- Autoencoders, variational autoencoders

INTRODUCTION AND MOTIVATION

Supervised learning

- Given a training set with pairs of inputs x and corresponding desired outputs y

$$\Omega_{\text{train}} = \{(x^{(1)}, y^{(1)}), \ldots, (x^{(m)}, y^{(m)})\}$$

- Create a function f that "approximates" this mapping

$$f(x) \approx y, \quad \forall (x, y) \in \Omega_{\text{train}}$$

- Hope that this generalises well to unseen examples, such that

$$f(x) = \hat{y} \approx y, \quad \forall (x, y) \in \Omega_{\text{test}}$$

  where Ω_{test} is a set of relevant unseen examples.
- Hope that this is also true for all unseen relevant examples.

Unsupervised learning

- In contrast with supervised learning, we have *no* labeled data points in unsupervised learning.
- Since there is no “ground truth”, there is no accuracy evaluation in the supervised sense.
- Applications
  - Data clustering
  - Anomaly detection
  - Signal generation
  - Signal compression

Semi-supervised learning

- We have *some* labeled data
- Usually a majority of unlabeled data
- Can be thought of as supervised learning extended to utilise unlabeled data
- Will not be covered today

Methods

What we *will* cover today

- K-means clustering (background)
- Principal component analysis (PCA) (background)
- t-SNE
- Autoencoders
- Variational autoencoders

What we *will not* cover today

- Independent component analysis (ICA)
- Matrix factorization and decomposition
- Expectation-maximization (EM) algorithm
- Generative-adverserial networks (GAN) (next lecture)

CLUSTERING

Data clustering

- Grouping together data based on some similarity metric
- Data points within the same group (cluster) will be more similar to each other than to data points outside the group
- Many different versions of clustering

Connectivity-based clustering

- Also called hierarchical clustering
- See figures for example with the L_2 distance metric measured from cluster centroides
- Different level thresholds yields different clusters

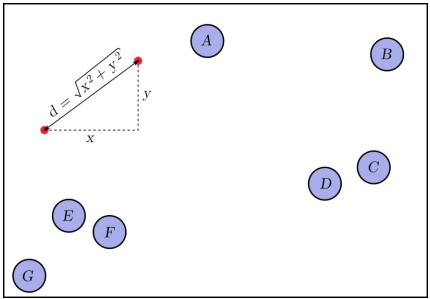


Figure 1: Raw data

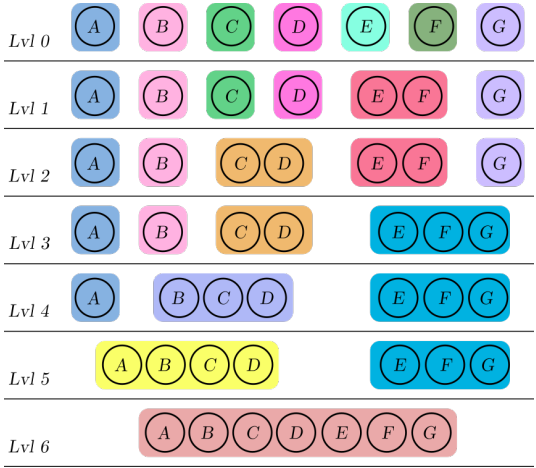


Figure 2: Bottom up (agglomerative) hierarchy of clusters

Graph clustering — cliques

- A clique is a set of nodes
- A node in a clique shares an edge with all other nodes in the clique
- Can have cliques of different sizes
- Useful in areas such as random fields

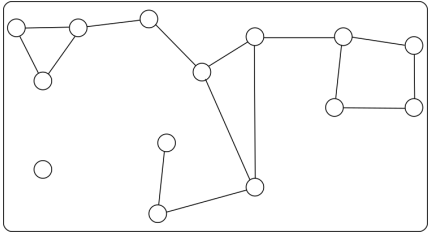


Figure 3: Undirected graph

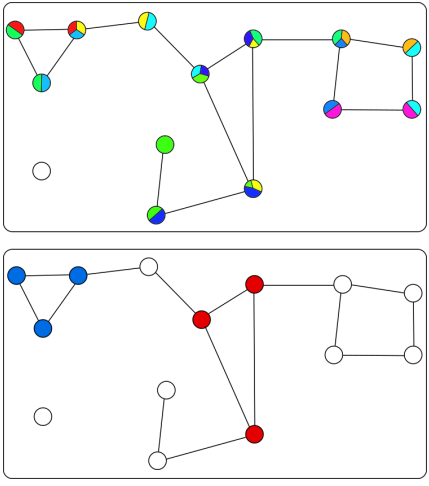


Figure 4: Top: Cliques with 2 members. Bottom: Cliques with 3 members. Nodes with multiple colors belong to more than one clique.

Centroid-based clustering

- Clusters are represented by a central vector
- Example: K-means clustering

K-means clustering

- Conseptually simple clustering algorithm
- We want to partition a set of data $\{x^{(1)}, x^{(2)}, \ldots, x^{(m)}\}$ into k clusters.
- $x^{(i)} \in \mathbb{R}^n, i = 1, \ldots, m$
- With some distance norm $||\cdot||$ the procedure is
  1. Initialize at random k cluster centroids (or means) $\mu_j \in \mathbb{R}^n, j = 1, \ldots, k$
  2. Repeat until convergence
     2.1 Assign every example $x^{(i)}, i = 1, \ldots, m$ with the label of the nearest cluster centroid
     $$c^{(i)} = \arg\min_j ||x^{(i)} - \mu_j||.$$
     2.2 Update the position of every centroid $\mu_j, j = 1, \ldots, k$ to the centroid of the cluster of points with its label
     $$\mu_j = \frac{\sum_{i=1}^m I[c^{(i)} = j] x^{(i)}}{\sum_{i=1}^m I[c^{(i)} = j]},$$
     where the *Iverson bracket* is defined as
     $$I[a = b] = \begin{cases} 1, & \text{if} \quad a = b, \\ 0, & \text{if} \quad a \neq b \end{cases}.$$

- Minimizes the objective function

$$J(c, \mu) = \sum_{i=1}^{m} ||x^{(i)} - \mu_{c^{(i)}}||$$

- Not guaranteed to find a *global minimum*
- Common to run the algorighm several times with different initializations, and then pick the run with the smallest value of J
- The k-means clustering algorithm partitions the feature space into *Voronoi cells*

K-means clustering — example

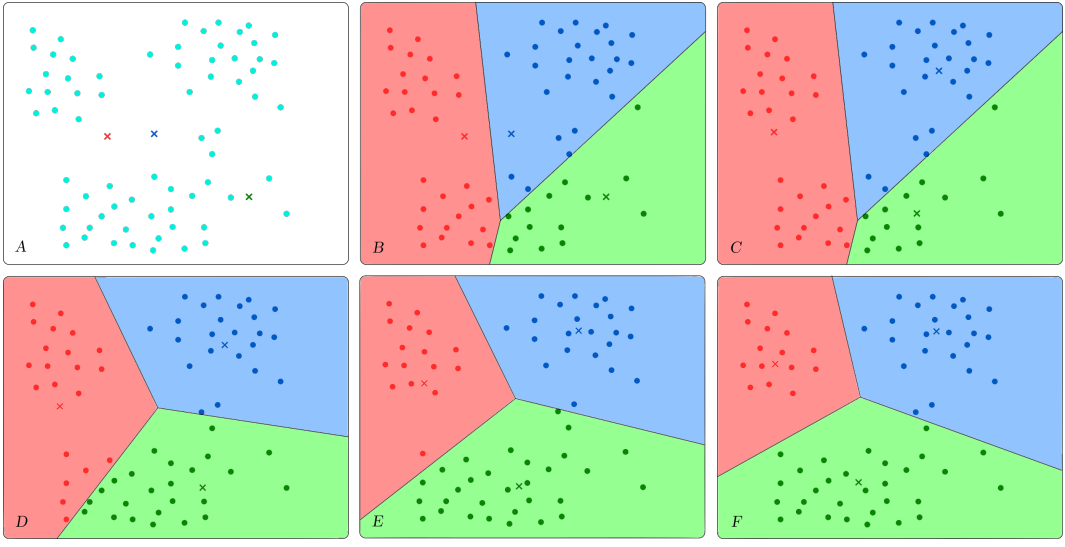


Figure 5: A: Initialize centroids. B: Assign points to clusters. C: Move centroids. D, E, F: Assign points to clusters and move centroids. No change after F (convergence).

PCA

Principal Component Analysis (PCA)

- Reducing the dimensionality of a dataset of correlated variables
- Retaining as much as possible of the variance present in the dataset

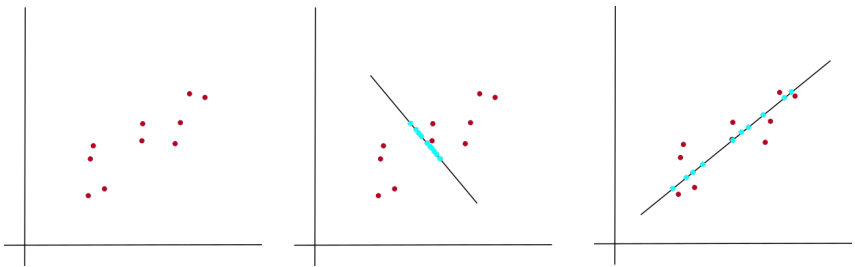


Figure 6: Representing 2D data as 1D

PCA — DERIVATION OUTLINE

- Let $X \in \mathrm{R}^{n_d}$ be a random vector
- We are looking for a set of *uncorrelated* variables Y_k which we will call the *principal components* of X
- The first component, Y_1, will account for most of the variance in X
- The second component, Y_2, will account for most of the variance in X, conditionied on being uncorrelated with Y_1
- The third component, Y_3, will account for most of the variance in X, conditioned on being uncorrelated with *both* Y_1 and Y_2
- We continue untill we have $n_p << n_d$ principal components that account for most of the variance in X

PCA — FIRST PRINCIPAL COMPONENT

- Let $Y_1 \in \mathbb{R}$ be some linear combination of the elements in X

$$Y_1 = \sum_{i=1}^{N_d} a_{1i} X_i = a_1^\intercal X,$$

- This random variable has variance

$$Var[Y_1] = Var[a_1^\intercal X] = a_1^\intercal \Sigma a_1.$$

- Here, Σ is the covariance matrix of X with elements

$$\Sigma_{ij} = Cov(X_i, X_j)$$

- We want to maximize the variance of Y_1
- In order to achieve finite solutions, we constrain the optimization on

$$a_1^\intercal a_1 = 1$$

- It turns out that, for $k = 1, \ldots, n_p$, a_k well be an eigenvector of Σ corresponding to the kth largest eigenvalue λ_k

- For a dataset with n_s samples $\{x_{i1}, \ldots, x_{in_s}\}$ for all features $i = 1, \ldots, n_d$, the elements in the covariance matrix can be estimated as

$$\hat{\Sigma}_{ij} = \frac{1}{n_s - 1} \sum_{q=1}^{n_s} (x_{iq} - \hat{\mu}_i)(x_{jq} - \hat{\mu}_j),$$

- Here $\hat{\mu}_i$ is the sample mean of the ith feature

$$\hat{\mu}_i = \frac{1}{n_s} \sum_{q=1}^{n_s} x_{iq}$$

- Arranginging the feature samples and sample means into vectors of size n_d

$$x_q = [x_{1q}, \ldots, x_{n_d q}]^\mathsf{T}$$
$$\hat{\mu} = [\hat{\mu}_1, \ldots, \hat{\mu}_{n_d}]^\mathsf{T}$$

- With this, the estimate of the covariance matrix can be written as

$$\hat{\Sigma} = \frac{1}{n_s - 1} \sum_{q=1}^{n_s} (x_q - \hat{\mu})(x_q - \hat{\mu})^\mathsf{T}.$$

PCA — Optimizing the variance

- We use the technique of *Lagrangian multipliers* to incorporate the unit length constraint
- This means that we are going to maximize the expression

$$J(a_1) = a_1^\intercal \Sigma a_1 - \lambda(a_1^\intercal a_1 - 1).$$

- Computing the gradient of J w.r.t. a_1, and setting it equal to zero, yields

$$\Sigma a_1 - \lambda a_1 = 0,$$

or

$$(\Sigma - \lambda I) a_1 = 0,$$

where I is the $n_d \times n_d$ identity matrix.

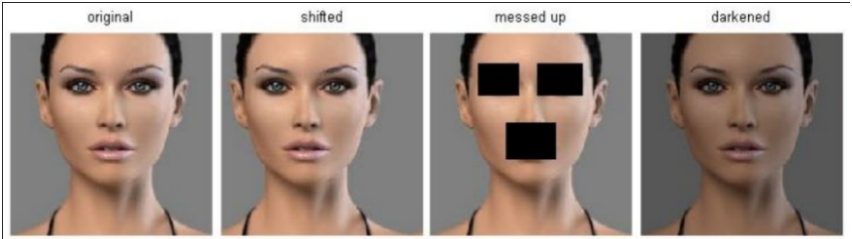
PCA — Optimizing the variance

- From our last expression
$$(\Sigma - \lambda I)a_1 = 0,$$
we see that λ is an eigenvalue of Σ, and a_1 is the corresponding eigenvector.
- Furthermore, λ is the largest eigenvalue
- This is because maximizing the variance subject to the constraint of unit length coefficients is equivalent to choosing the largest eigenvalue
$$\begin{aligned} a_1^\intercal \Sigma a_1 &= a_1^\intercal \lambda a_1 \\ &= \lambda a_1^\intercal a_1 \\ &= \lambda. \end{aligned}$$
- In general, the kth principal component of X is
$$a_k^\intercal X$$
where a_k is the eigenvector of the covariance matrix Σ of X, corresponding to the kth largest eigenvalue λ_k

PCA — Applications

- Dimensionality reduction
- Preprocessing in supervised learning: acts as a regularizer
- Noise reduction

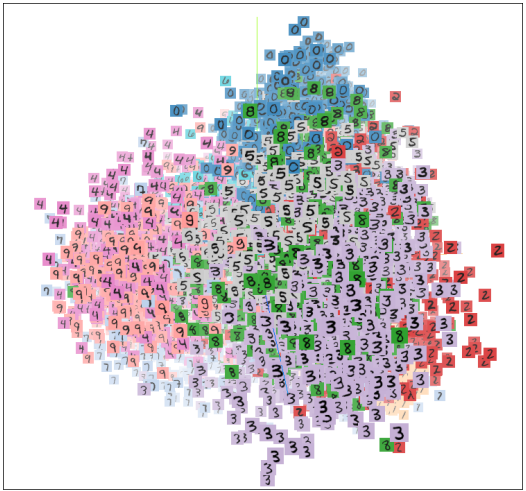
Problems with image data

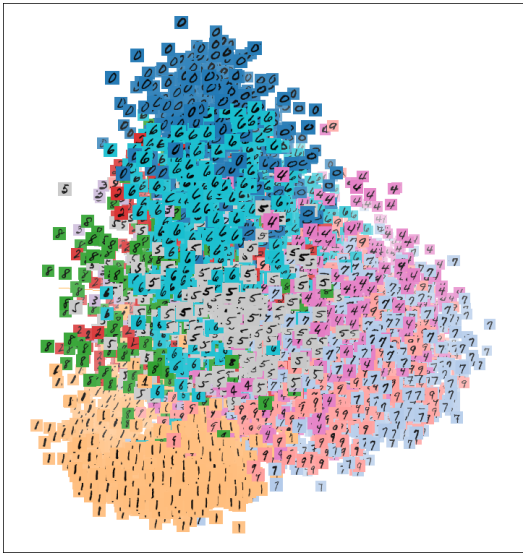


(all 3 images have same L2 distance to the one on the left)

MNIST clustering with PCA

Explains about 26% of the variance. Not very suited.





T-SNE

Stochastic Neighbour Embedding (SNE)

- Precursor to t-SNE (*t-distributed* Stochastic Neighbour Embedding)
- Introduced by Geoffrey Hinton and Sam Roweis in 2003 [1]
- A stochastic dimensionality reduction method
- Transforms high-dimensional (HD) data points to low-dimensional (LD) data points
- Aims to preserve neighbourhood relationship between data points
- Similar (close) HD points should also be similar (close) in the LD representation

[1] http://papers.nips.cc/paper/2276-stochastic-neighbor-embedding.pdf

SNE — overview

- The high-dimensional points have some dimension h
- The low-dimensional points have some desired predetermined dimension $l << h$
- For each point i, we are going to define two distributions:
  - $P_i(x_j)$: Describes the probability that point j is the "neighbour" of point i, given its location x_i
  - $Q_i(y_j)$: Describes the probability that point j is the "neighbour" of point i, given its location y_i
- We are then going to define a similarity measure between these distributions
- The low-dimensional representations will be altered such as to minimize this distribution similarity

- Let X be a h-dimensional random variable (RV) modelling a HD point
- Let S be a h-dimensional RV that is modelling a neighbour of X
- Given that $X = x_i$, we want the probability that S is a neighbour of X to be proportional to the Gaussian of the euclidian distance between the two

$$\Pr(S = s|X = x_i) = \frac{1}{c_i} \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left\{-\frac{||x - s||^2}{2\sigma_i^2}\right\}$$

  where c_i is a constant
- We define $\Pr(S = x_i|X = x_i) = 0$
- We also want it to be a probability, so if we sum over all possible neighbours $z \neq x_i$

$$\sum_{z \neq x_i} \Pr(S = z|X = x_i) = 1$$

- We end up with

$$\Pr(S = s|X = x_i) = \frac{\exp\left\{-\frac{||x_i - s||^2}{2\sigma_i^2}\right\}}{\sum_{z \neq x_i} \exp\left\{-\frac{||x_i - z||^2}{2\sigma_i^2}\right\}}$$

- The probability mass function that describes the probability that some neighbour S of X is located at s given that X is located at x_i is

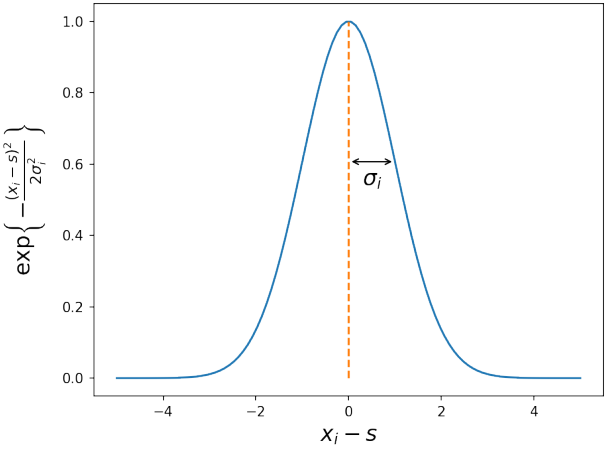
$$P_i(s) := \Pr(S = s | X = x_i)$$
$$:= \frac{\exp\left\{-\frac{||x_i - s||^2}{2\sigma_i^2}\right\}}{\sum_{z \neq x_i} \exp\left\{-\frac{||x_i - z||^2}{2\sigma_i^2}\right\}}$$

- Given a concrete set of points $\{x_1, x_2, \ldots, x_n\}$
- The probability that j is a neighbour of i, given that i is located at x_i is then

$$p_{j|i} := \frac{\exp\left\{-\frac{||x_i - x_j||^2}{2\sigma_i^2}\right\}}{\sum_{k \neq i} \exp\left\{-\frac{||x_i - x_k||^2}{2\sigma_i^2}\right\}}$$

SNE — SCALING PARAMETER

- The scaling parameter σ_x can be set manually
- We want a larger σ_x in sparse areas
- We want a smaller σ_x in dense areas



SNE — PERPLEXITY

- σ is often found with binary search such that the perplexity equals k, which is determined manually
- The perplexity of the distribution P_i is given by

$$Perp(P_i) = 2^{H(P_i)}$$

where the Shannon entropy is given by

$$H(P_i) = -\sum_j p_{j|i} \log_2 p_{j|i}$$

- Perplexity can be interpreted as a measure of how many neighbours we want to influence a point
- Typical values are between 5 and 50
- See e.g. https://distill.pub/2016/misread-tsne/ how to interpret t-SNE results

SNE — LOW-DIMENSION NEIGHBOUR PROBABILITY

- Let Y be a l-dimensional RV modelling a LD point
- Let T be a l-dimensional RV that is modelling a neighbour of Y
- Y are the lower-dimensional data points corresponding to X, so $l << h$
- Similarly to HD, we choose a Gaussian neighbourhood, but with fixed variance $\sigma^2 = 1/2$

$$\Pr(T = t|Y = y_i) = \frac{\exp\left\{-||y_i - t||^2\right\}}{\sum_{z \neq y_i} \exp\left\{-||y_i - z||^2\right\}}$$

SNE — low-dimension neighbour probability, notation

- For every HD point x_i, we have a corresponding LD point y_i
- The probability mass function that describes the probability that some neighbour T of Y is located at t given that Y is located at y_i is

$$\begin{aligned} Q_i(s) &:= \Pr(T = t | Y = y_i) \\ &:= \frac{\exp\left\{-||y - t||^2\right\}}{\sum_{z \neq x_i} \exp\left\{-||y_i - z||^2\right\}} \end{aligned}$$

- Given a concrete set of points $\{y_1, y_2, \ldots, y_n\}$
- The probability that j is a neighbour of i, given that i is located at y_i is then

$$q_{j|i} := \frac{\exp\left\{-||y_i - y_j||^2\right\}}{\sum_{k \neq i} \exp\left\{-||y_i - y_k||^2\right\}}$$

- The goal is to place y_i such that the LD distribution $q_{j|i}$ is similar to the HD distribution $p_{j|i}$
- We need a similarity metric, and a way to optimize it

Repetition: Kullback-Leibler divergence

- The Kullback-Liebler divergence over a discrete random variable X

$$D_{KL}(p_X||q_X) = \sum_x p_X(x) \log \frac{p_X(x)}{q_X(x)}$$

- Measures the distance between two probability distributions p_X and q_X over the same set of events, modeled with the random variable X.
- Expectation of logarithmic difference between p and q when expectation is taken w.r.t. p.
- Measures the amount of information that is lost when using q to approximate p.
- It is non-negative
- Zero for $p = q$
- Increasing for "increasing difference" between p and q.

SNE — DISTRIBUTION SIMILARITY MEASURE

- We want to measure the similarity between P_i and Q_i, for all points i
- This is done by summing the KL-divergence between the original (P_i) and the "induced" (Q_i) distributions over all points

$$\begin{aligned} C &= \sum_i D_{KL}(P_i||Q_i) \\ &= \sum_i \sum_j p_{j|i} \log \frac{p_{j|i}}{q_{j|i}} \end{aligned}$$

- Large cost of confusing a small distance in the high-dimensional space with a large distance in the low-dimensional space (small $p_{j|i}$ and large $q_{j|i}$)
- Larger cost of confusing a large distance in the high-dimensional space with a small distance in the low-dimensional space (large $p_{j|i}$ and small $q_{j|i}$)

SNE — discussion

- The cost can be minimized with stochastic gradient descent
- Note that we are minimizing w.r.t. the LD points $\{y_1, \ldots, y_n\}$ corresponding to the known HD points $\{x_1, \ldots, x_n\}$
- Keeps nearby points in HD nearby in LD
- Also keeps distant points in HD relatively far apart in LD
- Drawback: Can be difficult to optimize
- Drawback: Tendency to crowd LD representations at the center of the map ("crowding problem")

t-distributed Stochastic Neighbour Embedding (t-SNE)

- A variant of the SNE method
- Introduced by Laurens van der Maaten and Geoffrey Hinton in 2008 [2]
- An improvemet over SNE
  - Much easier to optimize
  - Significantly better visualization
- Two major differences between t-SNE and SNE
  - *Symmetric* Gaussian point similarity distribution for the HD data points
  - Student-t point similarity distribution for the LD map points

[2] https://lvdmaaten.github.io/publications/papers/JMLR_2008.pdf

Symmetric SNE

- Standard SNE use a sum over the KL-divergence between asymmetric conditional probability distributions

$$C = \sum_i D_{KL}(P_i||Q_i)$$

- Because of this, different types of errors in the pairwise distances in the map are weighted differently
- In particular
  - The cost of representing distant data points as close map points is smaller than
  - The cost of representing close data points as distant map points
- A symmetric cost could ease optimization, and leviate the crowding problem

Symmetric SNE

- In stead, we could use the KL-divergence between symmetric joint probability distributions

$$\begin{aligned} C &= D_{KL}(P||Q) \\ &= \sum_i \sum_j p_{ij} \log \frac{p_{ij}}{q_{ij}} \end{aligned}$$

- The joint probability p_{ij} over all points X_i and their neighbours X_j is

$$p_{ij} = \Pr(X_i = x_i, X_j = x_j)$$

- Again, we define $p_{ii} = 0$, and require that the sum over the entire possibility space (a point and its neighbours for all points) is 1
- With the same Gaussian neighbourhoods as in SNE, we get

$$p_{ij} = \frac{\exp\left\{-\frac{||x_i - x_j||^2}{2\sigma_i^2}\right\}}{\sum_k \sum_{l \neq k} \exp\left\{-\frac{||x_k - x_l||^2}{2\sigma_k^2}\right\}}$$

Symmetric SNE

- Similarly, for the LD points

$$
\begin{aligned}
q_{ij} &= \Pr(Y_i = y_i, Y_j = y_j) \\
&= \frac{\exp\left\{-\frac{||y_i - y_j||^2}{2\sigma_i^2}\right\}}{\sum_k \sum_{l \neq k} \exp\left\{-\frac{||y_k - y_l||^2}{2\sigma_k^2}\right\}}
\end{aligned}
$$

- Note that $p_{ij} = p_{ji}$ and $q_{ij} = q_{ji}$
- Note that this is *not* what is used in t-SNE, we will come back to that in two slides
- This is just motivation

- We suggested a symmetric, joint probability

$$p_{ij} = \frac{\exp\left\{-\frac{||x_i - x_j||^2}{2\sigma_i^2}\right\}}{\sum_k \sum_{l \neq k} \exp\left\{-\frac{||x_k - x_l||^2}{2\sigma_k^2}\right\}}$$

- The problem is that for an outlier x_i, $||x_i - x_j||$ will be very large (and p_{ij} very small) for all points
- The placement of the corresponding point y_i will have very little effect on the cost

$$C = \sum_i \sum_j p_{ij} \log \frac{p_{ij}}{q_{ij}}$$

- We can fix this by simply using our previous conditional probabilities as

$$p_{ij} = \frac{p_{j|i} + p_{i|j}}{2n}$$

  where n is the number of data points
- With this, we ensure that $\sum_j p_{ij} > 1/(2n)$ for all data points x_i
- Hence, all points x_i are guaranteed to make significant contributions to the cost

t-SNE — Crowding problem

- Standard SNE (and other similar methods) suffer from what is known as the *crowding problem*
- Too many map points are placed near the center of the map
- This can be leviated by forcing moderately distant data points to be placed far apart

- To mitigate the crowding problem, we want to give more weight to representing moderately distant data points as close map points
- The Student-t distribution with one degree of freedom is used

$$q_{ij} = \frac{\left(1 + ||y_i - y_j||^2\right)^{-1}}{\sum_{k \neq l} \left(1 + ||y_k - y_l||^2\right)^{-1}}$$

- Notice that it is symmetric $q_{ij} = q_{ji}$
- The Student-t distribution has a much heavier tail than the Gaussian distribution
- Moderate distances in the HD data space are then represented by larger distances in the LD map space

AUTOENCODERS

Autoencoders — introduction

- An autoencoder is a neural network which purpose is to discover interesting representations of data
- The idea is to create identity mappings, that is, functions f such that $f(x) \approx x$ for some input x
- It is able to discover interesting representations by enforcing constraints on the network
- The method requires no labeled data, and is therefore unsupervised

- An autoencoder consist of an encoder g and an decoder h
- The encoder maps the input x to some representation z

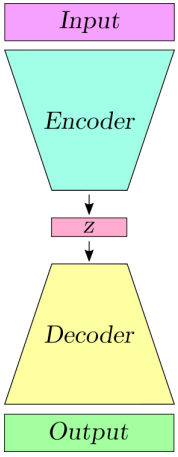
$$g : x \mapsto z$$

- The decoder maps this representation z to some output $\hat{x}$

$$h : z \mapsto \hat{x}$$

- We want to train the encoder and decoder such that

$$\begin{aligned} f(x) &= h(g(x)) \\ &= h(z) \\ &= \hat{x} \\ &\approx \hat{x} \end{aligned}$$



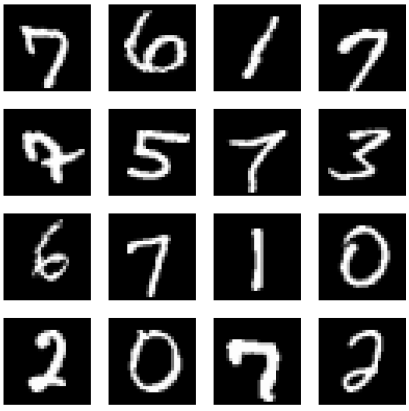
Autoencoders — variants

- Different network constraints leads to different representations z
- Compression autoencoder
  - If x has d_x dimensions and z has d_z dimensions, and $d_x > d_z$
  - Most common way of constraining the network
- Denoising autoencoder
  - Distorting the input x with some random noise
  - Leads to robust representations, resiliant to corrupted input
- Sparse autoencoder
  - z can actually have a greater dimension than x
  - Only allowing a subset of the hidden units to fire at the same time

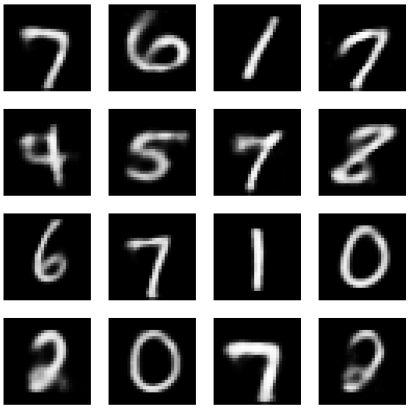
Compression autoencoder — MNIST example

- Encoder:
  - Input -> first hidden layer: fully connected 784 -> 128, relu
  - 1. hidden -> 2. hidden: fully connected, 128 -> 32, relu
- Decoder:
  - 2.hidden -> 3. hidden: fully connected 32 -> 128, relu
  - 3. hidden -> output: fully connected, 128 -> 784, sigmoid

Compression autoencoder — MNIST example



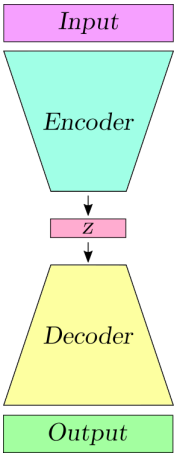
(a) Original



(b) Reconstructed

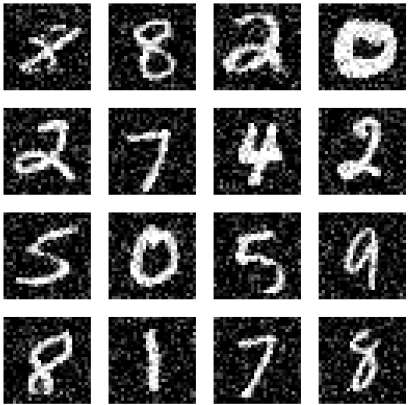
Denoising encoder

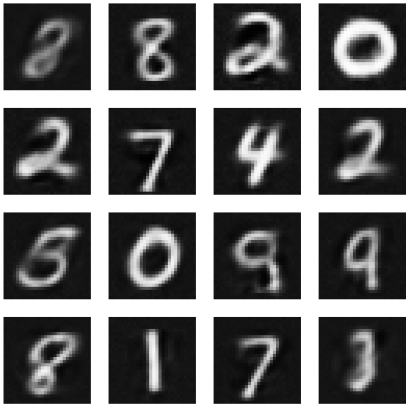
- Same set-up as in a compression autoencoder
- Add noise to the input
- Compare the reconstruction to the input without noise



Denoising autoencoder — MNIST example

Same setup as for the compression autoencoder. Zero mean Gaussian noise with standard deviation 0.1 is added to the input. The input values are clipped to lay in $[0, 1]$.





Sparse autoencoder

- We want to constrain the number of *active* nodes in the coding layer
- We can think of a node being active (or *firing*) if is
  - close to 1 for the sigmoid tanh activation functions
- We can think of a node being inactive
  - close to 0 for the sigmoid activation function
  - close to -1 for the tanh activation function
- We would like to constrain the nodes to be inactive most of the time

Sparse autoencoder

- Let $a_j^{[c]}(x^{(i)})$ be the activation in node j in the coding layer $[c]$ given an input $x^{(i)}$ to the network
- Then, activation for this node averaged over all m input examples is

$$\hat{\rho} = \frac{1}{m} \sum_{i=1}^{m} a_j^{[c]}(x^{(i)})$$

- We would like to limit this average activation by enforcing the constraint

$$\hat{\rho} = \rho$$

  for some predetermined *sparsity parameter* ρ
- Choosing a small ρ (e.g. 0.1) forces the activations to be small

Sparse autoencoder

- The way we enforce this constraint is to regularize the loss function

$$L = L_{\text{reconstruction}} + \beta L_{\text{sparsity}}$$

  with some regularization strength $\beta \in \mathbb{R}$.
- We are going to use the KL-divergence between the distributions p and q_j summed over the entire latent layer as our sparsity loss

$$L_{\text{sparsity}} = \sum_{j=1}^{n^{[c]}} D_{KL}(p||q_j)$$

  where $n^{[c]}$ is the number of nodes in layer $[c]$
- p will be a Bernoulli distribution with mean ρ for a node j
- q_j will be a Bernoulli distribution with mean $\hat{\rho}_j$ for a node j
- The Bernoulli distribution describes the probability of an event with two outcomes (e.g. coin toss)
- In our case p will represent a node being active with probability ρ, and q_j a node being active with probability $\hat{\rho}_j$

In this case, the KL divergence for a single node j is

$$\begin{aligned} D_{KL}(p||q_j) &= \sum p(x) \log \frac{p(x)}{q_j(x)} \\ &= \sum p(x) \log p(x) - \sum p(x) \log q_j(x) \end{aligned}$$

The support of the distributions is only two outcomes $x \in \{0, 1\}$, and the pmf is

$$p(x) = \begin{cases} (1-\rho), & x = 0 \quad \text{(the node is inactive)} \\ \rho, & x = 1 \quad \text{(the node is active)} \end{cases}$$

and conversely for $q_j(x)$. With this, our KL divergence is simply

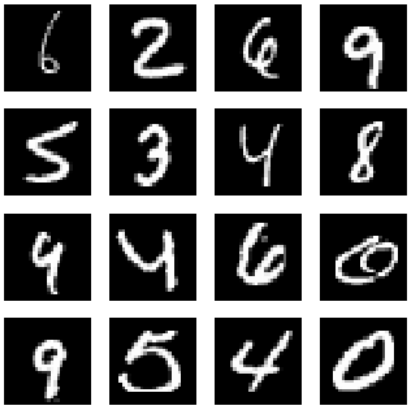
$$\begin{aligned} D_{KL}(p||q_j) &= \rho \log \rho + (1-\rho) \log(1-\rho) - [\rho \log \hat{\rho} + (1-\rho) \log(1-\hat{\rho}_j)] \\ &= \rho \log \frac{\rho}{\hat{\rho}_j} + (1-\rho) \log \frac{(1-\rho)}{(1-\hat{\rho}_j)}. \end{aligned}$$

- With this, we get our final loss

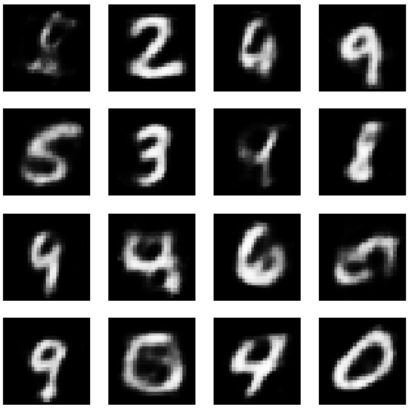
$$L = L_{\text{reconstruction}} + \beta \sum_{j=1}^{n^{[c]}} \rho \log \frac{\rho}{\hat{\rho}_j} + (1-\rho) \log \frac{(1-\rho)}{(1-\hat{\rho}_j)}$$

- Remember that $\hat{\rho}_j$ is the jth component of $\hat{\rho} = \frac{1}{m} \sum_{i=1}^{m} a_j^{[c]}(x^{(i)})$
- This means that we need to average over all examples to compute $\hat{\rho}$
- This means that we have to encode all said examples
- In practice, with batch optimization, we average over all examples in a batch

Sparse autoencoder — MNIST example



(a) Original



(b) Reconstructed

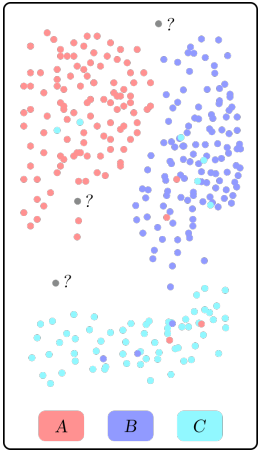
VARIATIONAL AUTOENCODERS

Introduction

- Popular method for signal generation (images, sound, language, etc.)
- Creating completely new signals
- Or altering existing data
- Especially powerful when you want to alter your data in a specific way, not just randomly

Problems with autoencoders for signal generation

- An autoencoder works great if you want to reconstruct a *replica* of the input
- Not well suited for generating new signal
- The reason for this is an “unintuitive” latent variable space
- The latent space might be discontinuous
- Random sampling from an “unseen” region of the latent space produces unpredictable results
- No reasonable way to interpolate between categories in the latent space

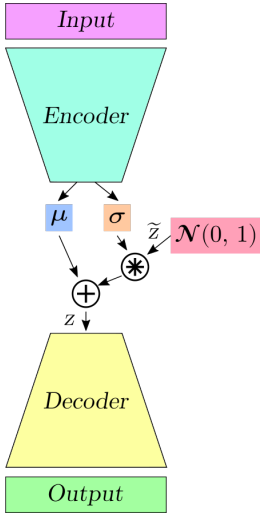


Variational autoencoders

- A variational autoencoder is designed to have a continuous latent space
- This makes them ideal for random sampling and interpolation
- It achieve this by forcing the encoder g to generate Gaussian representations, $z \sim \mathcal{N}(\mu, \sigma^2)$
- More precisely, for one input, the encoder generates a mean μ and a variance σ^2
- We then sample a zero-mean, unit-variance Gaussian $\tilde{z} \sim \mathcal{N}(0, 1)$
- Construct the input z to the decoder from this

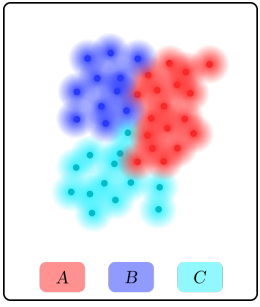
$$z = \mu + \tilde{z} \cdot \sigma$$

- With this, z is sampled from $q = \mathcal{N}(\mu, \sigma^2)$



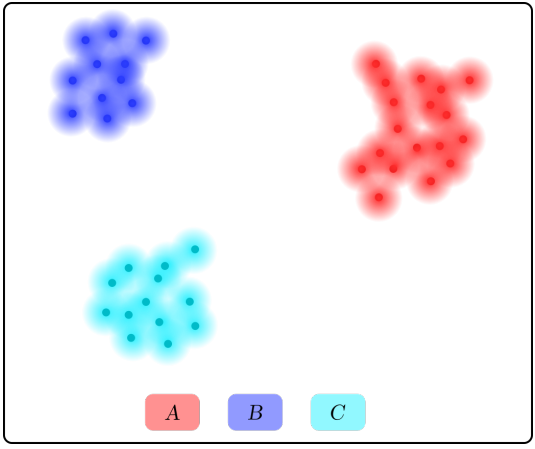
Intuition

- This is a stochastic sampling
- That is, we can sample different z from the same set of (μ, σ^2)
- The intuition is that the decoder "learns" that for a given input x:
  - the point z is important for reconstruction
  - but also a neighbourhood of z
- In this way, we have smoothed the latent space, at least locally



Problem

- No restriction on μ or σ^2
- Realisticly, clusters of different classes can be placed far apart
- Leaves “empty space” in between with unknown sampling features



Guiding the generative distribution

- We can guide the solutions by restricting the generative distribution q
- We do this by making it approximate some distribution p
- In that way, the latent vectors, even for different categories, will be relatively close
- The desired distribution used in variational autoencoders is the standard normal $p = \mathcal{N}(0, 1)$
- We use the familiar KL-divergence between the desired and the generated distribution as a regularizer in the loss function
- With this, the total loss for an example x_i is something like

$$L(x_i) = ||x^{(i)} - f(x^{(i)})|| + D_{KL}(p||q_{\mu_i,\sigma_i})$$

- That is, the sum of what we call the *reconstruction loss* and the *latent loss*
- The latent loss for a single input $x^{(i)}$ can be shown to be equal to

$$D_{KL}(p||q_{\mu_i,\sigma_i}) = \frac{1}{2}(\mu_i^2 + \sigma_i^2 - \log \sigma_i^2 - 1)$$

KL Divergence between Gaussian distributions

For reference, I will spend some slide deriving the KL Divergence between two Gaussian distributions $p = \mathcal{N}(\mu_p, \sigma_p^2)$ and $q = \mathcal{N}(\mu_q, \sigma_q^2)$.
We are going to derive it for the continuous case, where the KL-Divergence can be expressed as

$$\begin{aligned} D_{KL}(p||q) &= \int p(x) \log \frac{p(x)}{q(x)} \, \mathrm{d}x \\ &= \int p(x) \log p(x) \, \mathrm{d}x - \int p(x) \log q(x) \, \mathrm{d}x \end{aligned}$$

We will derive the two terms in the last line seperately

First, for the first term

$$\begin{aligned}\int p(x)\log p(x)\,\mathrm{d}x &= \int p(x)\log\left[(2\pi\sigma_p^2)^{-\frac{1}{2}}\exp\left\{-\frac{(x-\mu_p)^2}{2\sigma_p^2}\right\}\right]\mathrm{d}x \\ &= -\frac{1}{2}\log(2\pi\sigma_p^2)\int p(x)\,\mathrm{d}x - \frac{1}{2}\int p(x)\frac{(x-\mu_p)^2}{\sigma_p^2}\,\mathrm{d}x \qquad (1)\\ &= -\frac{1}{2}\log(2\pi\sigma_p^2) - \frac{1}{2\sigma_p^2}\int p(x)(x^2-2x\mu_p+\mu_p^2)\,\mathrm{d}x.\end{aligned}$$

Similarly, for the second term

$$\int p(x)\log q(x)\,\mathrm{d}x = -\frac{1}{2}\log(2\pi\sigma_q^2) - \frac{1}{2\sigma_q^2}\int p(x)(x^2-2x\mu_q+\mu_q^2)\,\mathrm{d}x. \qquad (2)$$

Remember that for a random variable X with pdf f, the expectation is given by

$$E[X] = \int f(x)x \, \mathrm{d}x.$$

Also, we have

$$\begin{aligned} E[X^2] &= \int f(x)x^2 \, \mathrm{d}x \\ &= Var[X] + E[X]^2 \end{aligned}$$

For the integral in eq. (1), we then get

$$\begin{aligned} \frac{1}{2\sigma_p^2} \int p(x)(x^2 - 2x\mu_p + \mu_p^2) \, \mathrm{d}x &= \frac{1}{2\sigma_p^2}[(\sigma_p^2 + \mu_p^2) - 2\mu_p^2 + \mu_p^2] \\ &= \frac{1}{2}. \end{aligned} \tag{3}$$

The integral in eq. (2) is similar,

$$\begin{aligned}\frac{1}{2\sigma_q^2}\int p(x)(x^2 - 2x\mu_q + \mu_q^2)\,\mathrm{d}x &= \frac{1}{2\sigma_q^2}[(\sigma_p^2 + \mu_p^2) - 2\mu_p\mu_q + \mu_q^2] \\ &= \frac{\sigma_p^2 + (\mu_p - \mu_q)^2}{2\sigma_q^2}.\end{aligned} \tag{4}$$

Finishing up, using eq. (1) and eq. (2) via eq. (3) and eq. (4), we finally get

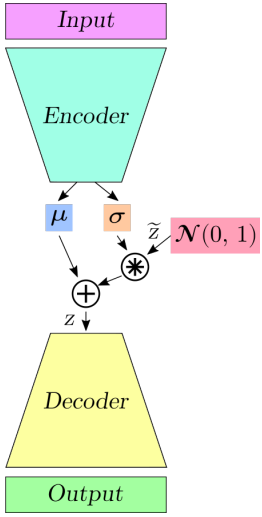
$$\begin{aligned} D_{KL}(p||q) &= \int p(x) \log \frac{p(x)}{q(x)} \, \mathrm{d}x \\ &= \int p(x) \log p(x) \, \mathrm{d}x - \int p(x) \log q(x) \, \mathrm{d}x \\ &= -\frac{1}{2} \log(2\pi\sigma_p^2) - \frac{1}{2} + \frac{1}{2} \log(2\pi\sigma_q^2) + \frac{\sigma_p^2 + (\mu_p - \mu_q)^2}{2\sigma_q^2} \\ &= \frac{1}{2} \left[\log \frac{\sigma_q^2}{\sigma_p^2} + \frac{\sigma_p^2 + (\mu_p - \mu_q)^2}{\sigma_q^2} - 1 \right] \end{aligned} \tag{5}$$

When, as in our case $p = \mathcal{N}(\mu, \sigma)$ and $q = \mathcal{N}(0, 1)$, we get

$$D_{KL}(p||q) = \frac{1}{2} \left[\mu^2 + \sigma^2 - \log \sigma^2 - 1 \right].$$

Generate new signals

- With a trained variational autoencoder $f = h \circ g$ you can generate new signals
- Sample $z \sim \mathcal{N}_{n^{[c]}}(0, 1)$, where $n^{[c]}$ is the number of nodes in the coding layer
- Feed z into the trained decoder h
- $h(z)$ should now be a randomly generated sample from the training distribution

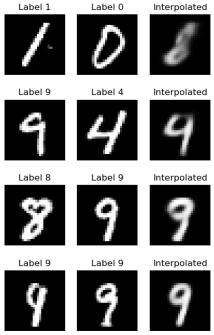


Generate new signals — interpolation

- Say you want to generate a signal c that is an interpolation between two signals a and b
- First, train a variational autoencoder $f = h \circ g$ on the desired distribution
- Compute mean vectors μ_a and μ_b from encodings $g(a)$ and $g(b)$
- Compute the average of the two mean vectors

$$\mu_c = \frac{1}{2}(\mu_a + \mu_b)$$

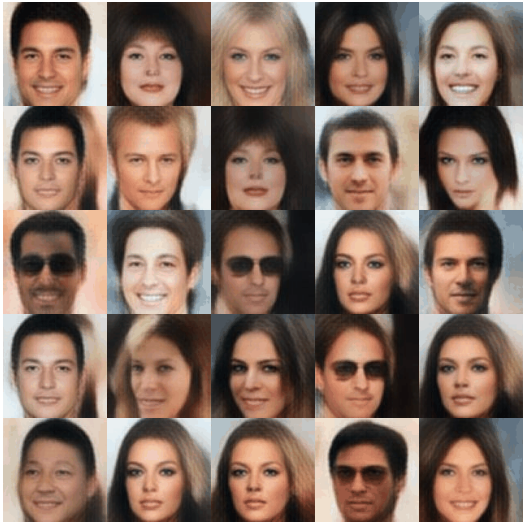
- Then, set the latent variable $z = \mu_c$
- $c = h(z)$ should then be an interpolation between a and b



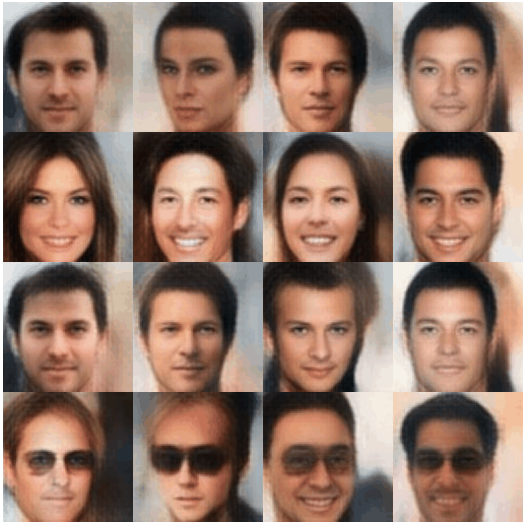
Generate new signals — adding features

- Say you want to add a feature of a signal a to the signal b
- You can do this by finding a signal c that is equal to a, except for the specific feature you want
- You can then subtract the latent variable of c from the latent variable of a, and add it to the latent variable of b
- Then you simply decode the new latent variable
- Example: “Face with glasses = face + (face with glasses — face)”
- See examples on the next slides

Variational autoencoders — Interpolation and combination example



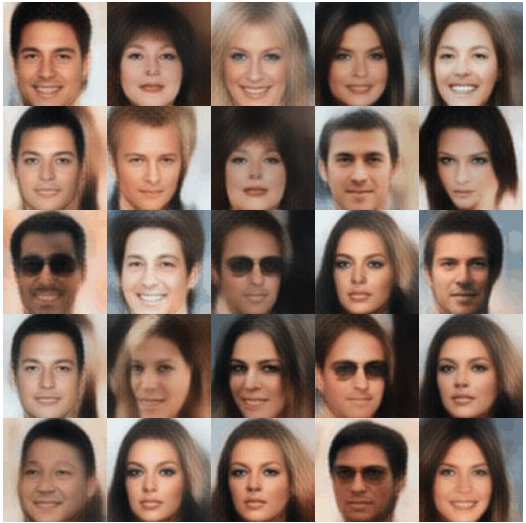
(a) Interpolation between genders



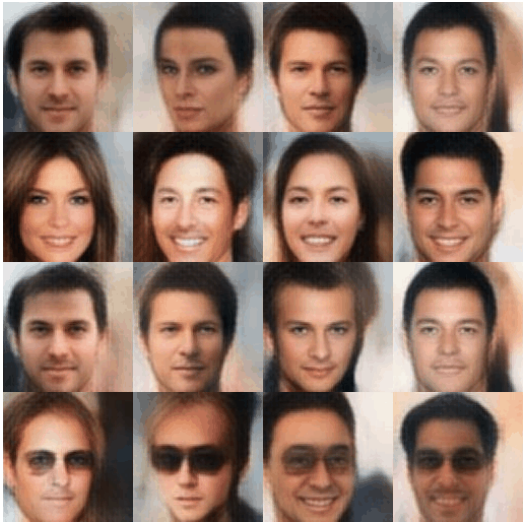
(b) Add or remove facial features

Figure 11: Source: Deep Feature Consistent Variational Autoencoder, https://houxianxu.github.io/assets/project/dfcvae

Variational autoencoders — Interpolation and combination example



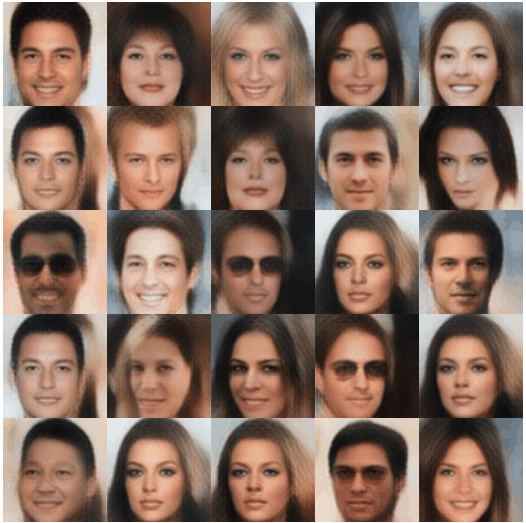
(a) Interpolation between genders



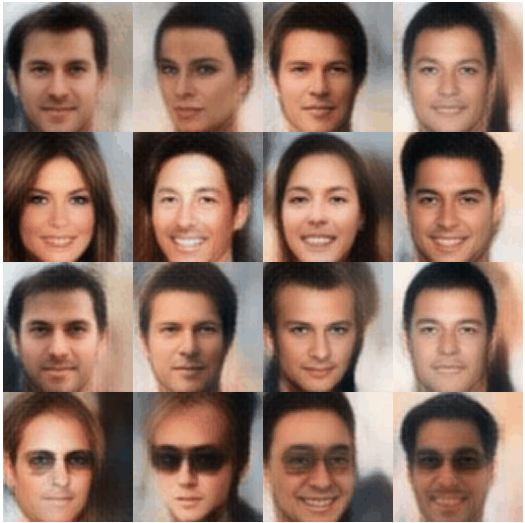
(b) Add or remove facial features

Figure 12: Source: Deep Feature Consistent Variational Autoencoder, https://houxianxu.github.io/assets/project/dfcvae

Variational autoencoders — Interpolation and combination example



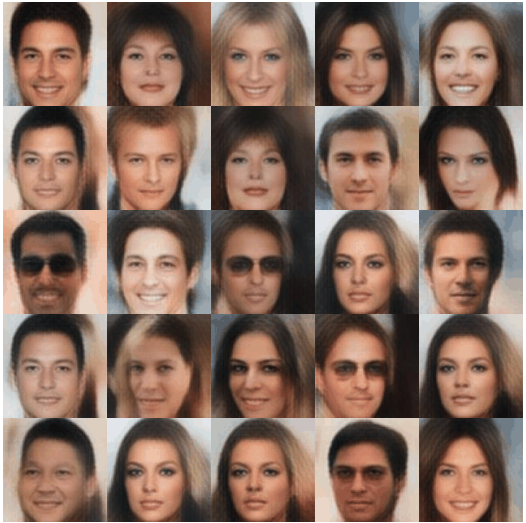
(a) Interpolation between genders



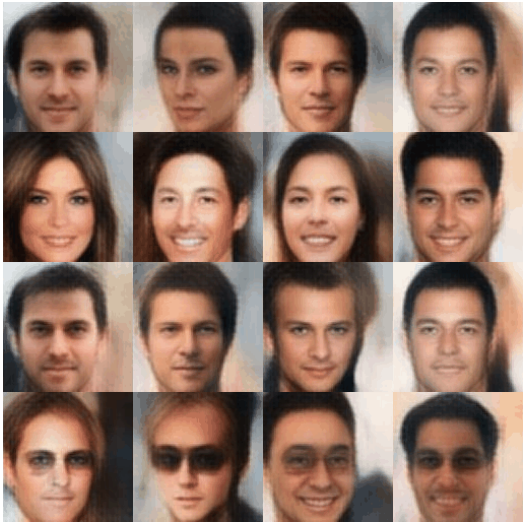
(b) Add or remove facial features

Figure 13: Source: Deep Feature Consistent Variational Autoencoder, https://houxianxu.github.io/assets/project/dfcvae

Variational autoencoders — Interpolation and combination example



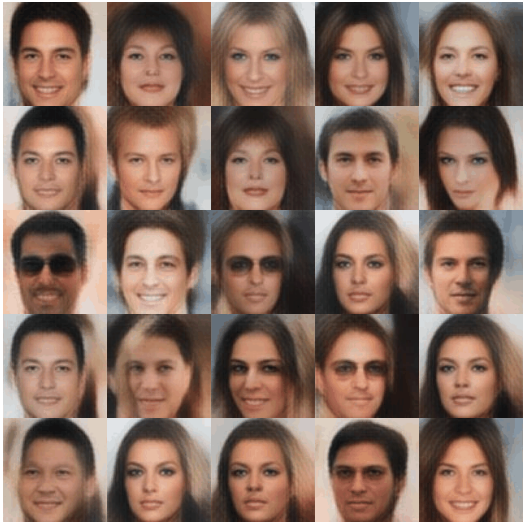
(a) Interpolation between genders



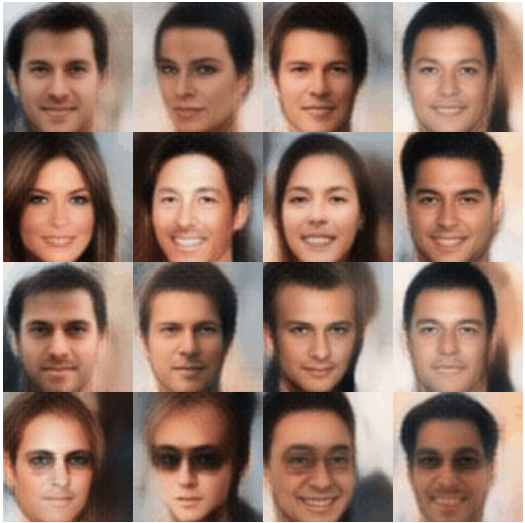
(b) Add or remove facial features

Figure 14: Source: Deep Feature Consistent Variational Autoencoder, https://houxianxu.github.io/assets/project/dfcvae

Variational autoencoders — Interpolation and combination example



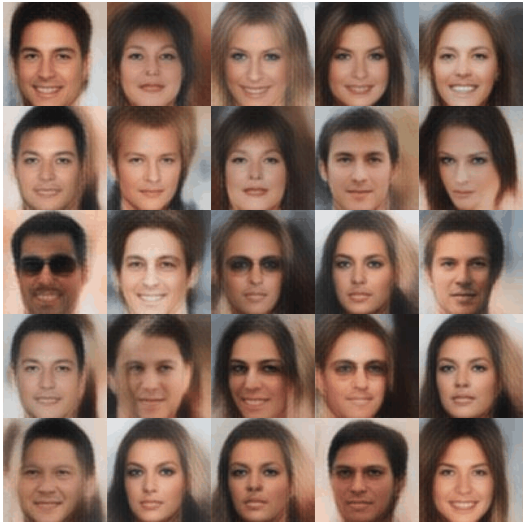
(a) Interpolation between genders



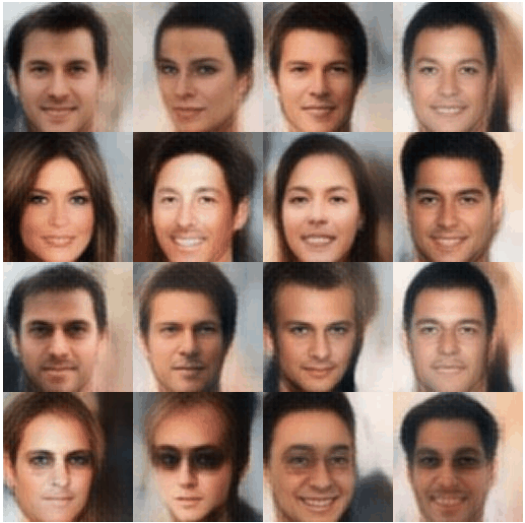
(b) Add or remove facial features

Figure 15: Source: Deep Feature Consistent Variational Autoencoder, https://houxianxu.github.io/assets/project/dfcvae

Variational autoencoders — Interpolation and combination example



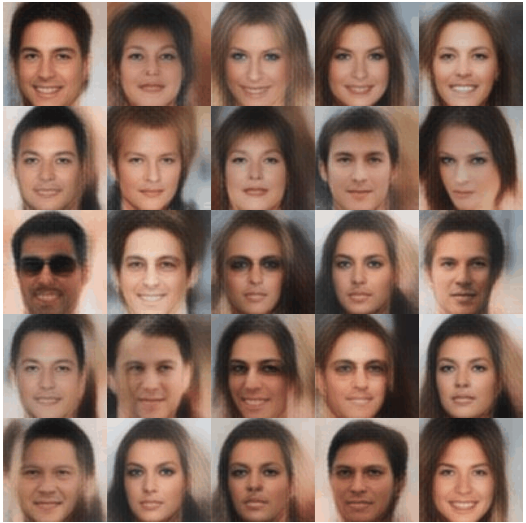
(a) Interpolation between genders



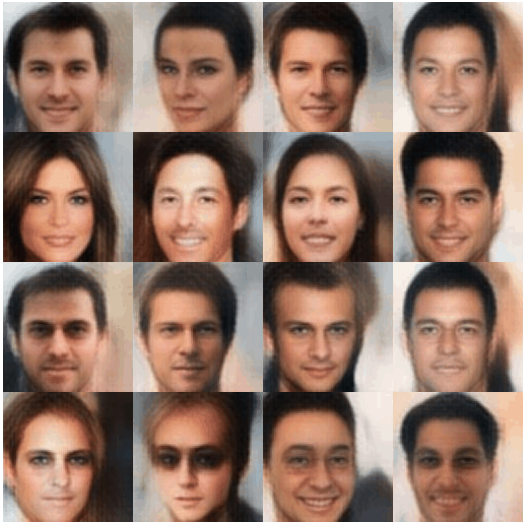
(b) Add or remove facial features

Figure 16: Source: Deep Feature Consistent Variational Autoencoder, https://houxianxu.github.io/assets/project/dfcvae

Variational autoencoders — Interpolation and combination example



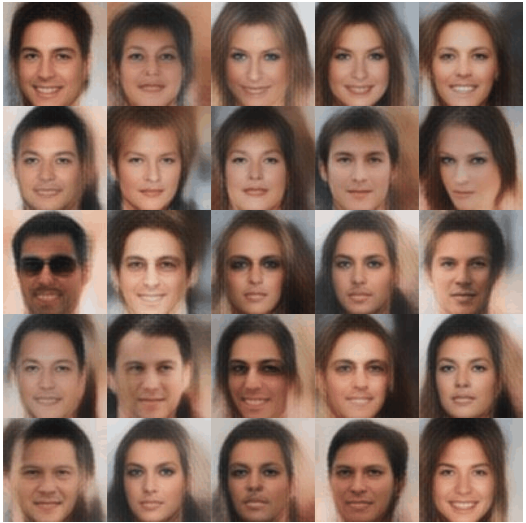
(a) Interpolation between genders



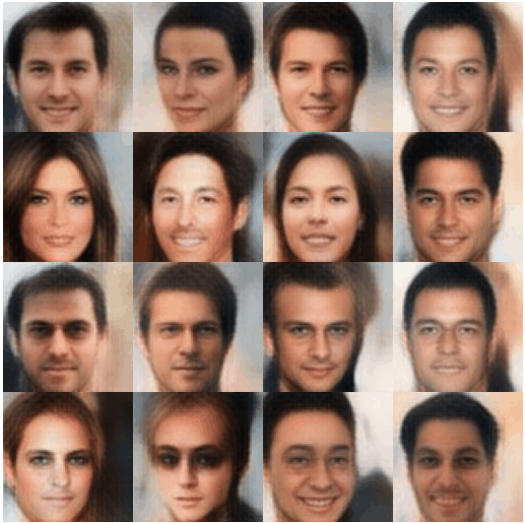
(b) Add or remove facial features

Figure 17: Source: Deep Feature Consistent Variational Autoencoder, https://houxianxu.github.io/assets/project/dfcvae

Variational autoencoders — Interpolation and combination example



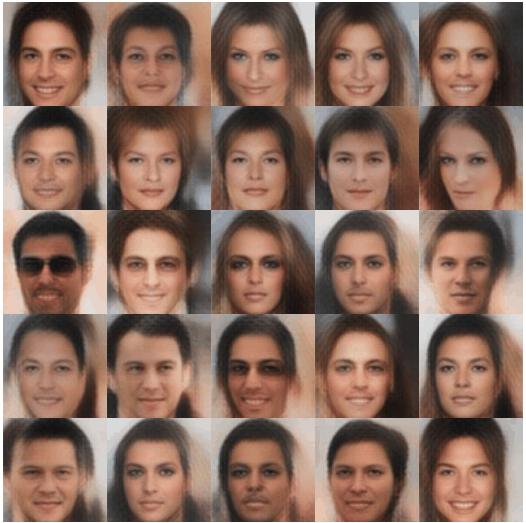
(a) Interpolation between genders



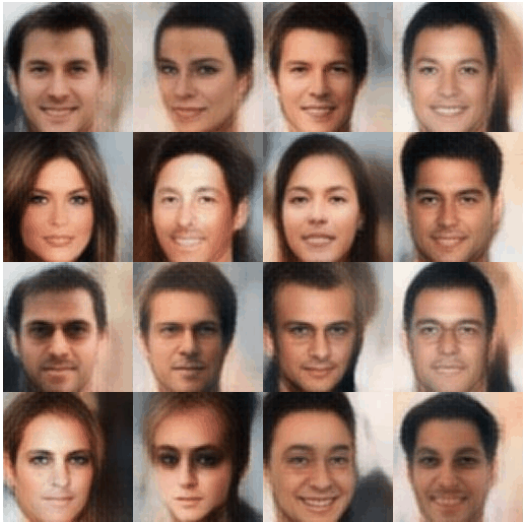
(b) Add or remove facial features

Figure 18: Source: Deep Feature Consistent Variational Autoencoder, https://houxianxu.github.io/assets/project/dfcvae

Variational autoencoders — Interpolation and combination example



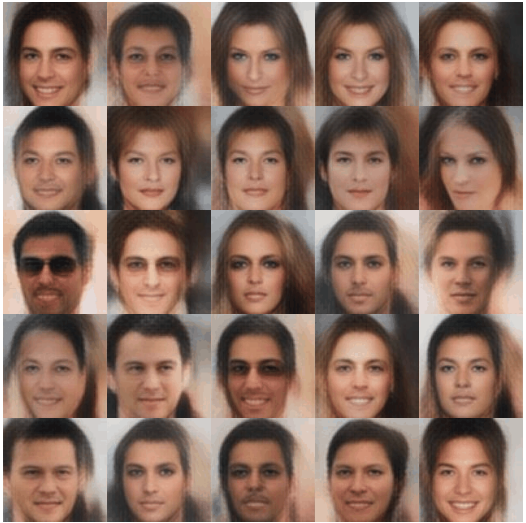
(a) Interpolation between genders



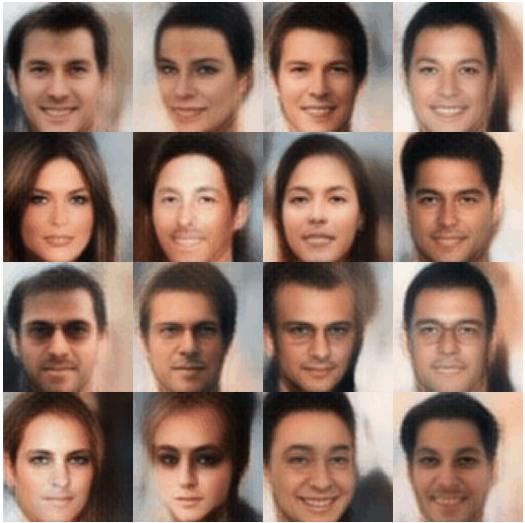
(b) Add or remove facial features

Figure 19: Source: Deep Feature Consistent Variational Autoencoder, https://houxianxu.github.io/assets/project/dfcvae

Variational autoencoders — Interpolation and combination example



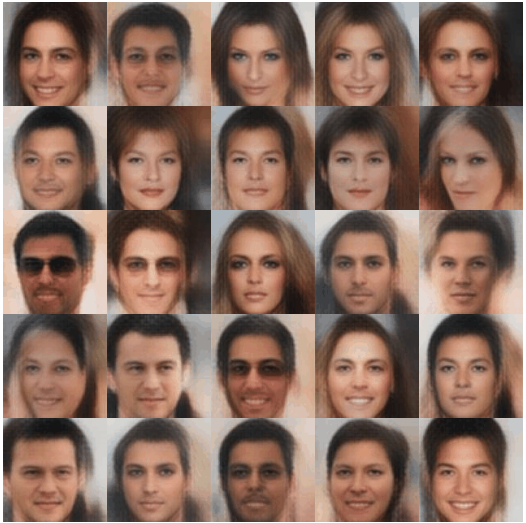
(a) Interpolation between genders



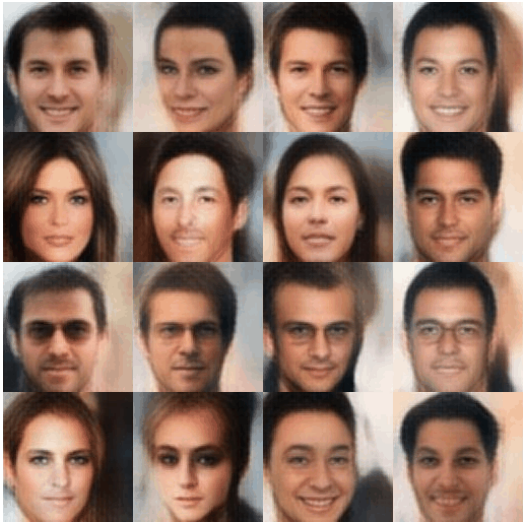
(b) Add or remove facial features

Figure 20: Source: Deep Feature Consistent Variational Autoencoder, https://houxianxu.github.io/assets/project/dfcvae

Variational autoencoders — Interpolation and combination example



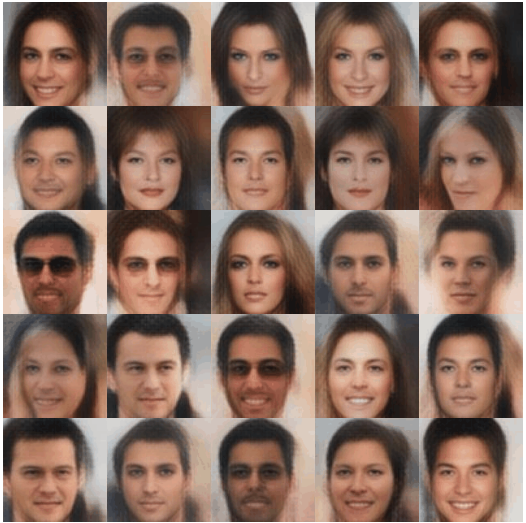
(a) Interpolation between genders



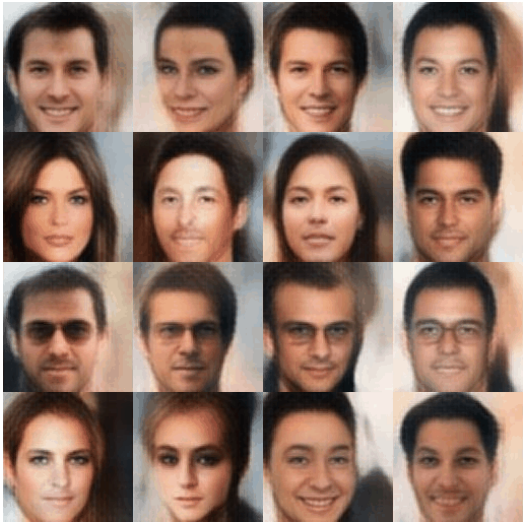
(b) Add or remove facial features

Figure 21: Source: Deep Feature Consistent Variational Autoencoder, https://houxianxu.github.io/assets/project/dfcvae

Variational autoencoders — Interpolation and combination example



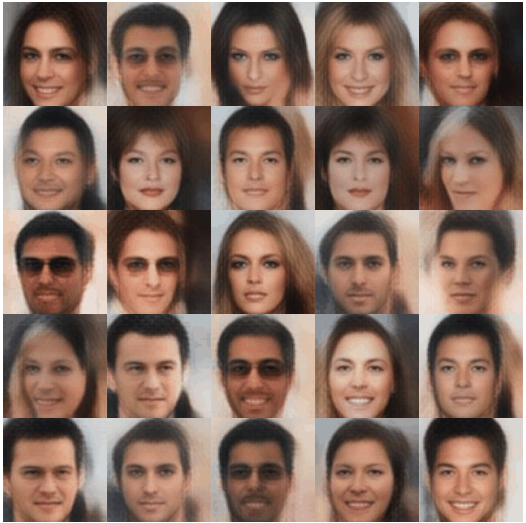
(a) Interpolation between genders



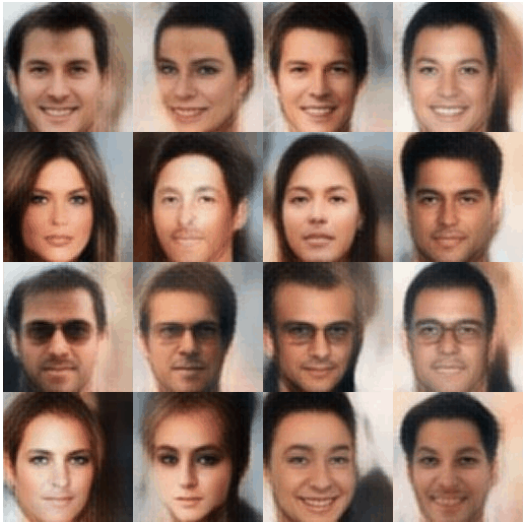
(b) Add or remove facial features

Figure 22: Source: Deep Feature Consistent Variational Autoencoder, https://houxianxu.github.io/assets/project/dfcvae

Variational autoencoders — Interpolation and combination example



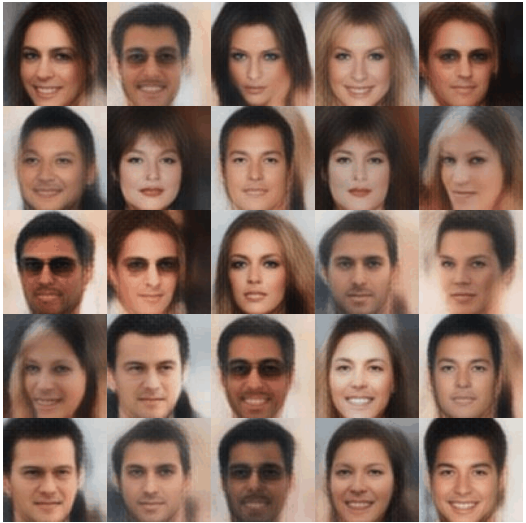
(a) Interpolation between genders



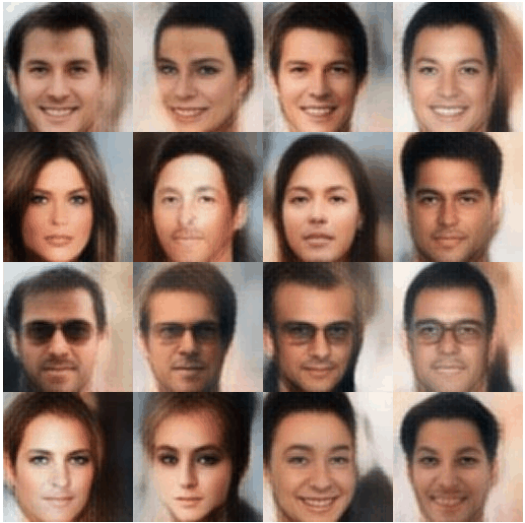
(b) Add or remove facial features

Figure 23: Source: Deep Feature Consistent Variational Autoencoder, https://houxianxu.github.io/assets/project/dfcvae

Variational autoencoders — Interpolation and combination example



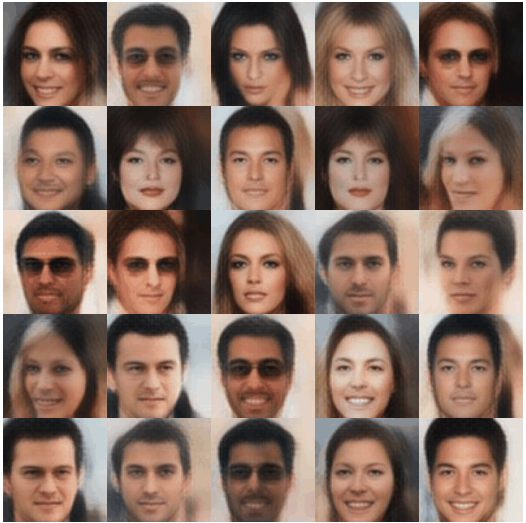
(a) Interpolation between genders



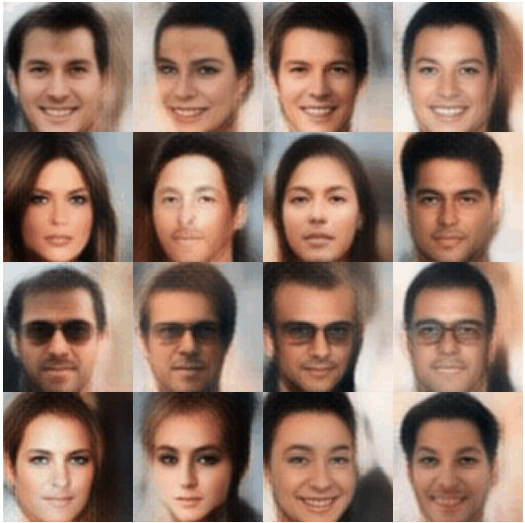
(b) Add or remove facial features

Figure 24: Source: Deep Feature Consistent Variational Autoencoder, https://houxianxu.github.io/assets/project/dfcvae

Variational autoencoders — Interpolation and combination example



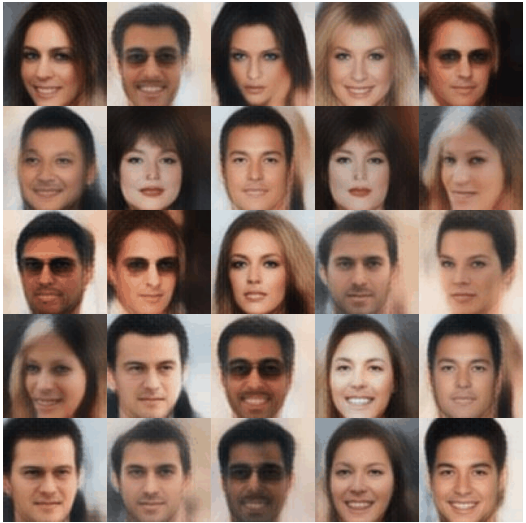
(a) Interpolation between genders



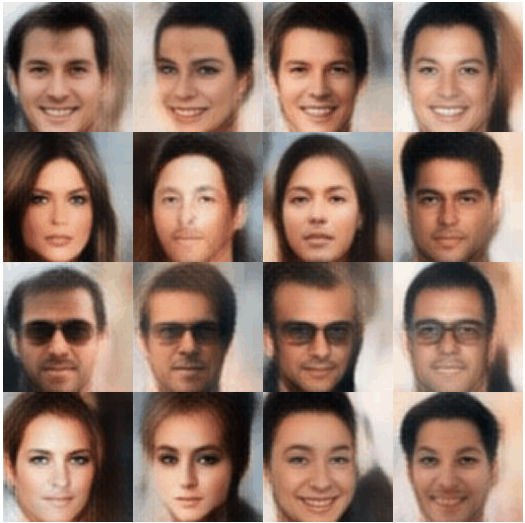
(b) Add or remove facial features

Figure 25: Source: Deep Feature Consistent Variational Autoencoder, https://houxianxu.github.io/assets/project/dfcvae

Variational autoencoders — Interpolation and combination example



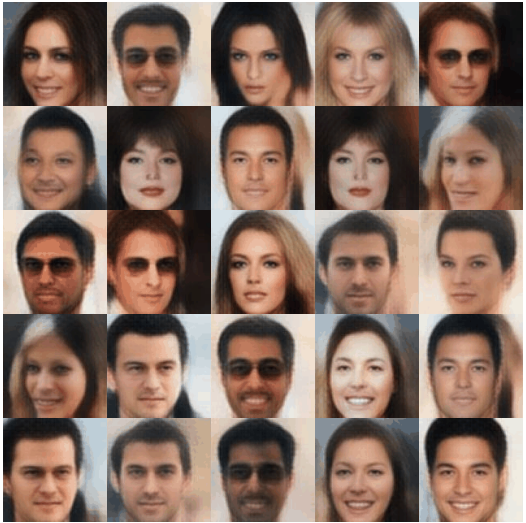
(a) Interpolation between genders



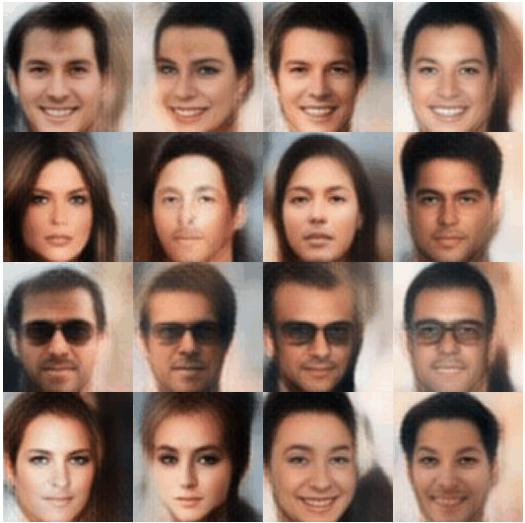
(b) Add or remove facial features

Figure 26: Source: Deep Feature Consistent Variational Autoencoder, https://houxianxu.github.io/assets/project/dfcvae

Variational autoencoders — Interpolation and combination example



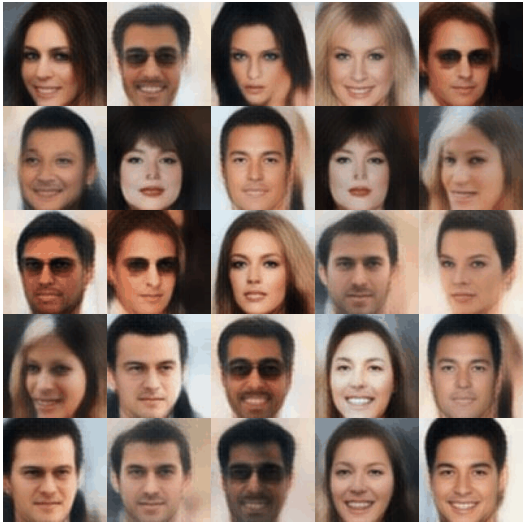
(a) Interpolation between genders



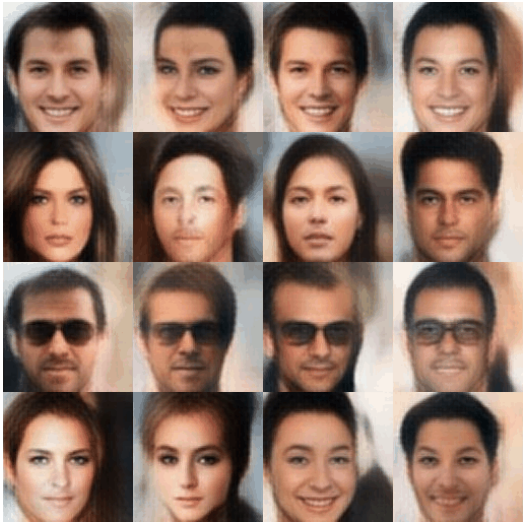
(b) Add or remove facial features

Figure 27: Source: Deep Feature Consistent Variational Autoencoder, https://houxianxu.github.io/assets/project/dfcvae

Variational autoencoders — Interpolation and combination example



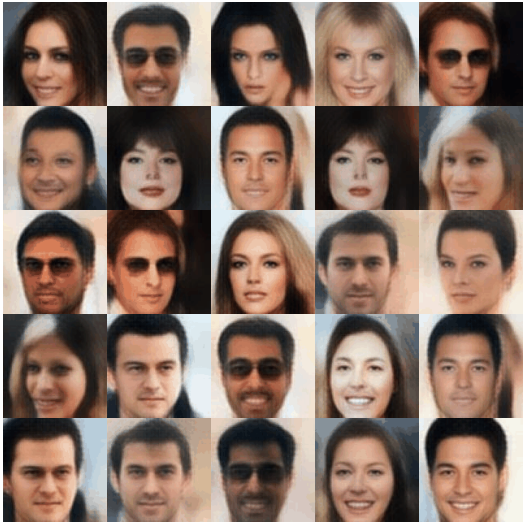
(a) Interpolation between genders



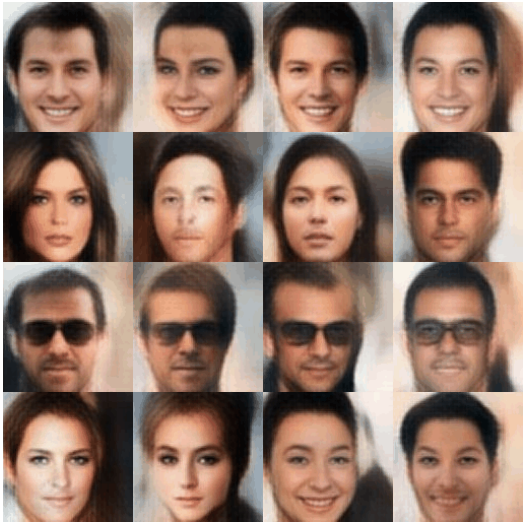
(b) Add or remove facial features

Figure 28: Source: Deep Feature Consistent Variational Autoencoder, https://houxianxu.github.io/assets/project/dfcvae

Variational autoencoders — Interpolation and combination example



(a) Interpolation between genders



(b) Add or remove facial features

Figure 29: Source: Deep Feature Consistent Variational Autoencoder, https://houxianxu.github.io/assets/project/dfcvae

Questions?