Deep Learning in Omics

Session 3: Stacked Denoising Autoencoders

FRC-EVL

Curso Extensión Universitaria - UB, 2019

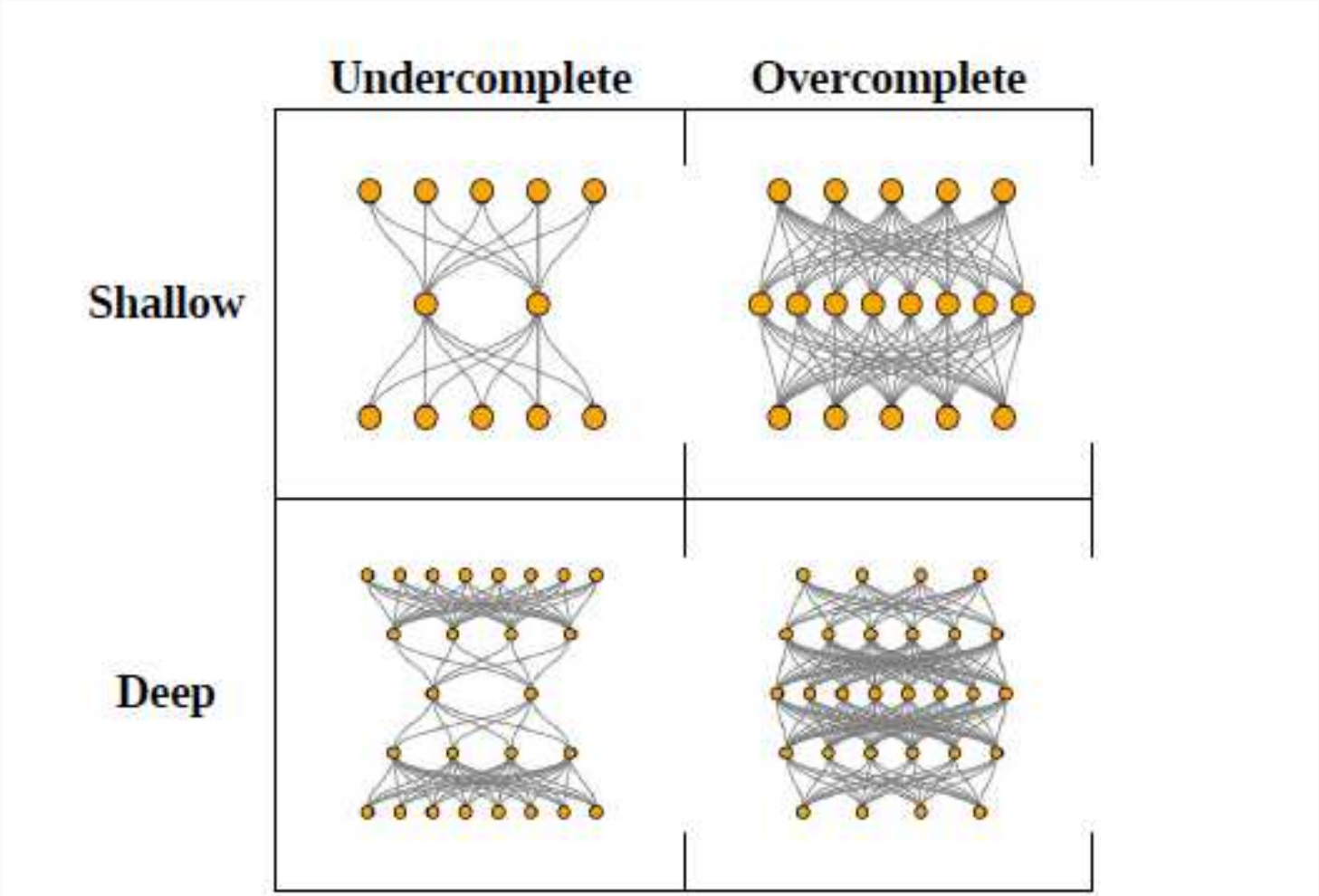
Table of Contents

1. Auto-encoders

Auto-encoders

- Multilayer Perceptron (MLP) is a feedforward neural network that maps the input to the output. A MLP is composed of nodes at multiple layers, including the input, output, and one or more hidden layers. Each layer in a MLP is fully connected with the next layer. In the hidden layers, each node is operated with a nonlinear activation function.
- An autoencoder is another type of neural networks that helps learning efficient codings of input data.
- With a primary goal of learning a compressed and distributed representation (i.e. encoding) of the input data, an autoencoder can thus be used for dimensionality reduction.
- In an auto-encoder, the output of the final hidden layer can be treated as a compressed representation of the input, if the hidden layers have fewer nodes than the input or output layers.

Auto-encoders classification

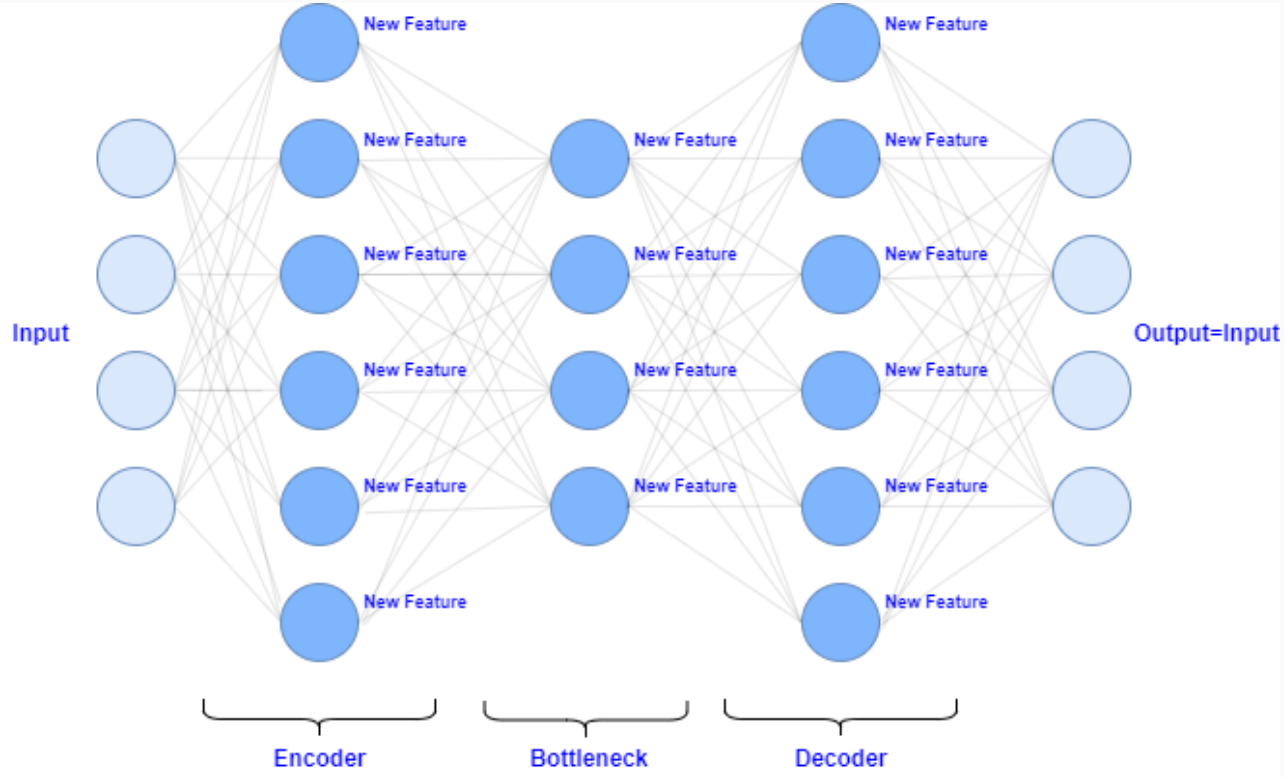


Auto-encoders

- An auto-encoder differs from a MLP in many ways. For example, the output layer of an auto-encoder has the same number of nodes as in the input layer.
- While an MLP can be learned to predict some target value y given the input x, an auto-encoder is trained to reconstruct its original input x by generating a reconstructed input $\hat{x}$ through optimizing its objective function.
- For an auto-encoder, the model tries to reproduce the provided input data x by using supervised learning, where the difference between the original input x and reconstructed input $\hat{x}$ is minimized.
- Backpropagation is also appropriate for training an auto-encoder.

Autoencoder in Representation Learning

The outputs of neurons in an Autoencoder are new representation of input data. While the standard role of Autoencoders is to compress data, in Representation Learning we normally use more neurons in hidden layers (in encoder, bottleneck and decoder layers) than number of features in input. With this unsupervised setup we can use not only training but also testing data to create new representation of features! We learn an Autoencoder to output the input data, and then we use neuron outputs (after application of activation function) as our new features.



2. Denoising auto-encoders

Denoising auto-encoders

- To build robust models from high-dimensional data, a denoising auto-encoder has been developed as an extension of a classical auto-encoder.
- The main goal of such a denoising auto-encoder is to separate signals from noises, which will allow the model to robustly reconstruct the output from partially destroyed input.
- Specifically, the corruption process of a denoising auto-encoder can be conducted in the following four steps.
    1. A process q is performed to corrupt the input X is corrupted.
    2. The corrupted input is mapped to Y via process f_θ .
    3. A process g_θ is conducted to reconstruct Y and generate the reconstruction of Z.
    4. The reconstruction error is measured by a loss function $L(X, Z)$, which will be used for backpropogation.

Denoising auto-encoders

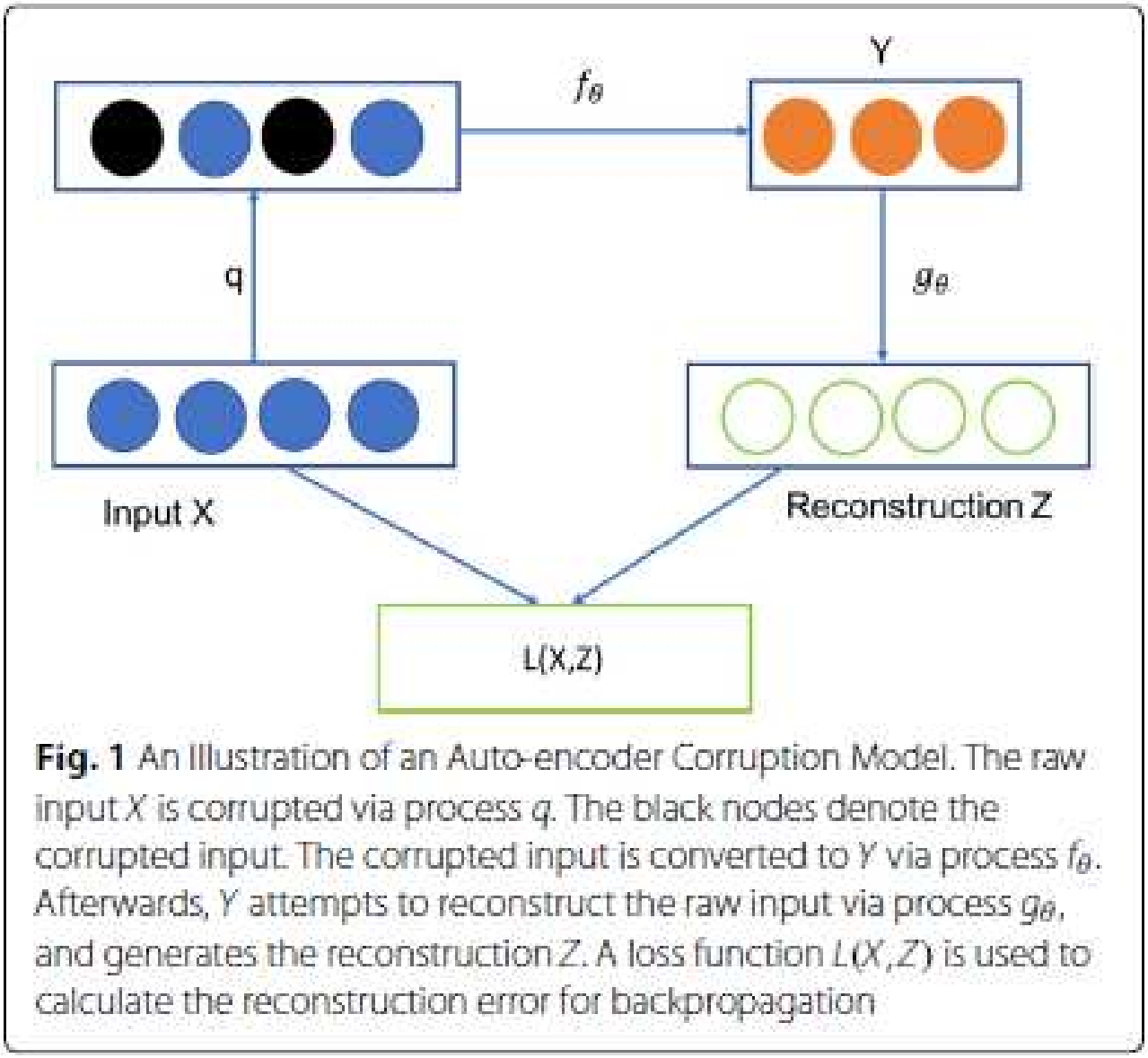


Fig. 1 An Illustration of an Auto-encoder Corruption Model. The raw input X is corrupted via process q. The black nodes denote the corrupted input. The corrupted input is converted to Y via process f_θ. Afterwards, Y attempts to reconstruct the raw input via process g_θ, and generates the reconstruction Z. A loss function $L(X,Z)$ is used to calculate the reconstruction error for backpropagation

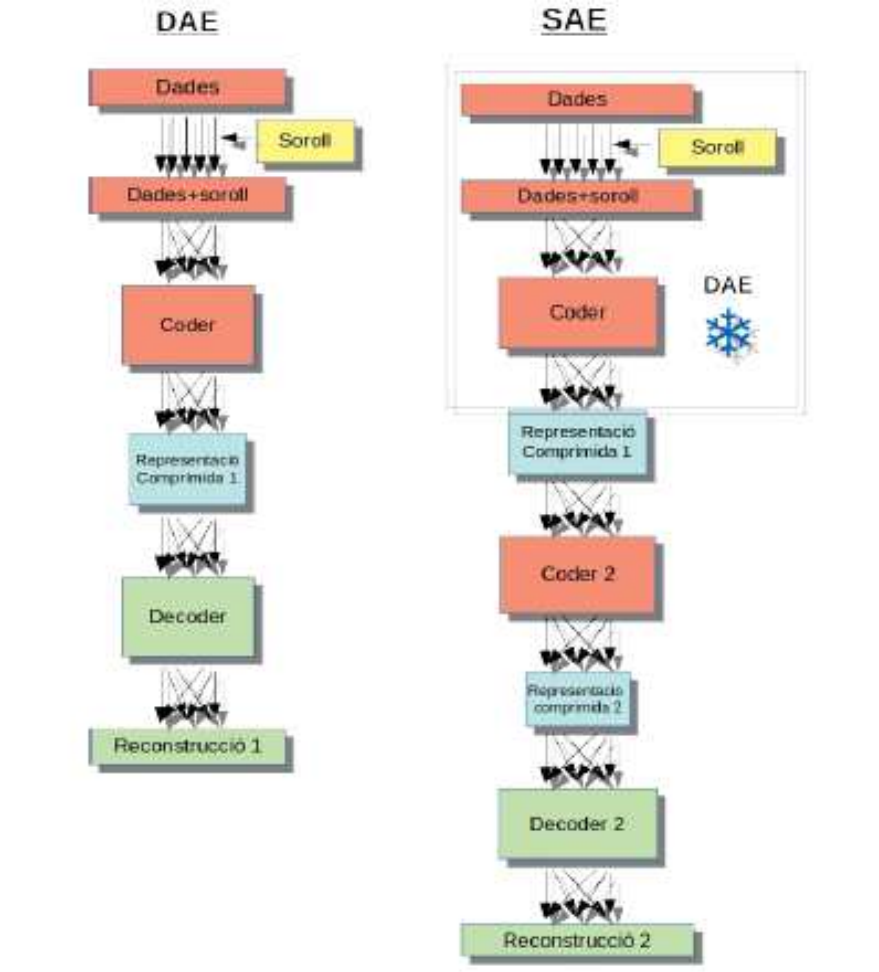
3. Stacked Denoising auto-encoders

Stacked Denoising auto-encoders

- The performance of a deep MLP is not good if we directly optimize a supervised objective function using algorithms like gradient descent with randomly initialized parameters.
- Denoising auto-encoders can be stacked as building blocks for constructing deep networks such as MLPs.
- A better MLP can be constructed by applying a local unsupervised learning to pre-train each layer in turn, and produce a useful higher-level representation from the lower-level one using the output from the previous layer.

Stacked Denoising auto-encoders

The SAE uses the encoder of pre-trained DAE model. In one first phase the weights of the encoder of the DAE are "frozen" to avoid the change during the training of the SAE model. In the second phase of "fine tunning" they are partially "un-frozen" to finish the adjustment of the weight of the complete model.



4. MLP-SAE in practice

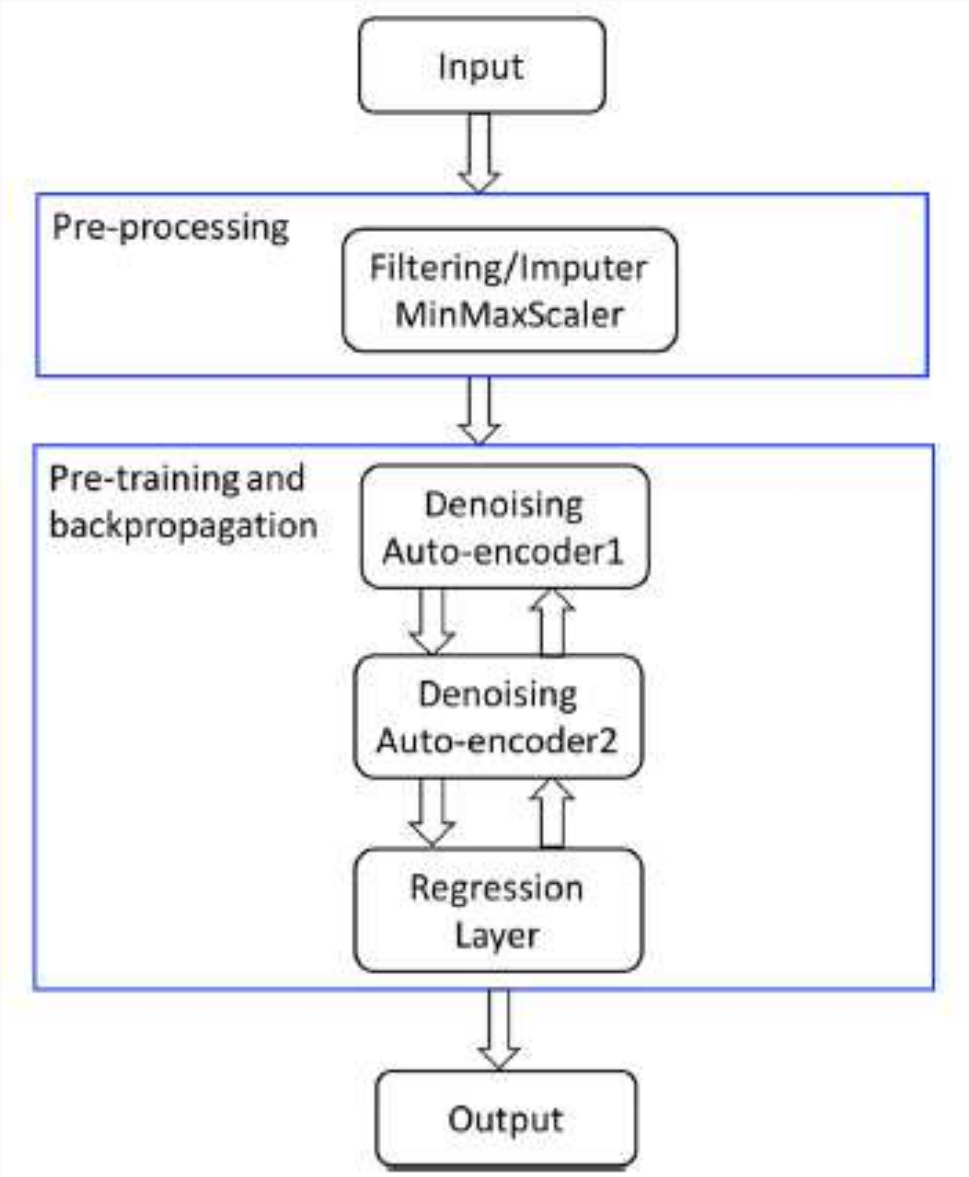
Genetic variants

- Studies have shown that genetic variants are associated with not only phenotypic traits of many kinds, but also linked with molecular traits such as gene expression.
- Therefore, assessing the effect of genetic variation on gene expression will improve our knowledge in understanding how genetic variation leads to phenotypic variation with regard to an organism's development, growth and survival.
- Expression QTL (eQTL) has been widely performed to study the influence of genetic variants on gene expression, where gene expression is considered as a quantitative trait.

MLP-SAE model

- To build a predictive model for estimating gene expression from genetic variation, we construct a deep denoising auto-encoder model utilizing the Multilayer Perceptron and Stacked Denoising Auto-Encoder (MLP-SAE).
- The proposed MLP-SAE model is composed of four blocks, one input, one output, and two hidden blocks including two auto encoders.
- The input layer takes input as SNP genotypes from yeast, with preprocessing conducted before feeding into the model.
- The output layer of the model is a regression model which generates the output as the predicted gene expression values.
- Stacked denoising auto-encoders are used as the hidden blocks of the model.
- The MLP-SAE model is trained and optimized by a backpropagation algorithm. The first training step is based on training the auto encoder with a stochastic gradient descent algorithm and the second training step utilizes the two auto-encoders as two hidden layers and training them with the multilayer perceptron.

MLP-SAE model



Data collection and pre-processing

Xie R, Wen J, Quitadamo A, Cheng J, Shi X. A deep auto-encoder model for gene expression prediction. BMC Genomics [Internet]. 2017 Nov 17 ;18(S9):845. Available from: https://bmcgenomics.biomedcentral.com/articles/10.1186/s12864-017-4226-0

Authors collect a widely-used yeast data set, with 2956 SNPs genotyped and the expression of 7085 genes measured in 112 samples which are crosses of the BY4716 and RM11- 1a strains.
They then remove missing values in the gene expression quantifications, resulted in the expression profiles of 6611 genes. They preprocess the SNP genotype file by conducting imputing and scaling.