# Notes on Autoencoders

compiled by D. Gueorguiev 3/25/24

An autoencoder is a specific type of a neural network, which is mainly designed to encode the input into a compressed and meaningful representation, and then decode it back such that the reconstructed input is as similar as possible to the original one. The purpose of autoencoders is learning in an unsupervised manner an “informative” representation of the data that can be used for various tasks such as clustering.

The problem is defined as to learn the functions (encoder) and (decoder) that satisfy:

(1)

where E is the expectation over the distributuion of x

Literature

[Autoencoders, Dor Bank et al, 2021](https://github.com/dimitarpg13/transformers_intro/blob/main/articles_and_books/Autoencoders.pdf)

[Autoencoders, Unsupervised Learning, and Deep Architectures, Pierre Baldi, 2012](https://github.com/dimitarpg13/deep_learning_and_neural_networks/blob/main/literature/articles/Autoencoders_Unsupervised_Learning_and_Deep_Architectures_Baldi_2012a.pdf)